Journal of Empirical Finance 78 (2024) 101517

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Non-standard errors in asset pricing: Mind your sorts☆

Amar Soebhag [a,b], Bart Van Vliet [a,c], Patrick Verwijmeren [a,d,*]

[a] Erasmus School of Economics, The Netherlands
[b] Robeco Quantitative Investing, The Netherlands
[c] Northern Trust Asset Management, The Netherlands
[d] University of Melbourne, Australia

ARTICLE INFO

JEL classification:
G11
G12
G15

Keywords:
Non-standard errors
Portfolio construction
Factor investing
Equity factors
Asset pricing models

ABSTRACT

Non-standard errors capture variation due to differences in research design choices. We document large variation in design choices in the context of asset pricing factor models and find that the average ratio of the non-standard error to the standard error across factors exceeds one. Using NAN breakpoints instead of NYSE breakpoints improves the average Sharpe ratios the most, from 0.46 to 0.63. Other important design choices relate to excluding microcaps, industry-adjusting, and the rebalancing frequency, which highlights the need for researchers to clearly describe and motivate these choices.

1. Introduction

Characteristic-based portfolio sorting is a widely used procedure in modern empirical finance. Researchers deploy the procedure to test theories in asset pricing, to study a wide range of pricing anomalies, and to identify profitable investment strategies. Based on this procedure, the academic literature in finance documents a range of factors that appear relevant for the cross-section of equity returns, known as the "factor zoo" (Cochrane, 2011). As researchers face a number of design choices when engaging in portfolio sorting, the exact construction procedure is not uniform across studies. Potentially, the differential design choices lead to considerable variation in outcomes. Menkveld et al. (2023) refer to such variation in outcomes due to choices in the evidence-generating process as non-standard errors (NSEs). In this paper, we study the extent to which the differential design choices in portfolio sorting matter for factors, factor models, and NSEs.

We first describe the choices made in a range of factor models and establish differences across at least six choices. These choices are (1) 70/30 or 80/20 breakpoints, (2) NYSE or NYSE-AMEX-Nasdaq (NAN) breakpoints, (3) 2×3 or $2 \times 3 \times 3$ sorting, (4) rebalancing frequency, (5) including or excluding financial firms, and (6) including or excluding firms with a negative book equity value. For other choices, such as whether to value-weight portfolio returns, factor models are uniform. In general, higher uniformity facilitates a comparison across factor model in terms of selecting a particular model and interpreting presented results. We then also consider choices made frequently in the wider asset pricing literature on portfolio sorting: (7) imposing a price filter or not, (8)

☆ We thank an anonymous referee, Guido Baltussen, Pedro Barroso, Victor DeMiguel, David Feldman, Amit Goyal, Paul Fontanier, Kewei Hou, Peter Koudijs, Frederik Muskens, Dimitris Papadimitriou, Esad Smaljbegovic, Sjoerd van Bekkum, Remco Zwinkels, seminar participants at Robeco, and conference participants at the 2022 Finance Symposium, 2022 Frontiers of Factor Investing conference, and 2023 Portuguese Finance Conference for fruitful discussions and feedback. We thank Wouter Lapre for excellent research assistance. The views expressed in this paper are not necessarily shared by Robeco Institutional Asset Management nor Northern Trust Asset Management.

* Correspondence to: Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3000 DR, The Netherlands.
E-mail addresses: soebhag@ese.eur.nl (A. Soebhag), b.p.vanvliet@ese.eur.nl (B. Van Vliet), verwijmeren@ese.eur.nl (P. Verwijmeren).

https://doi.org/10.1016/j.jempfin.2024.101517
Received 6 April 2023; Received in revised form 15 May 2024; Accepted 30 May 2024
Available online 4 June 2024

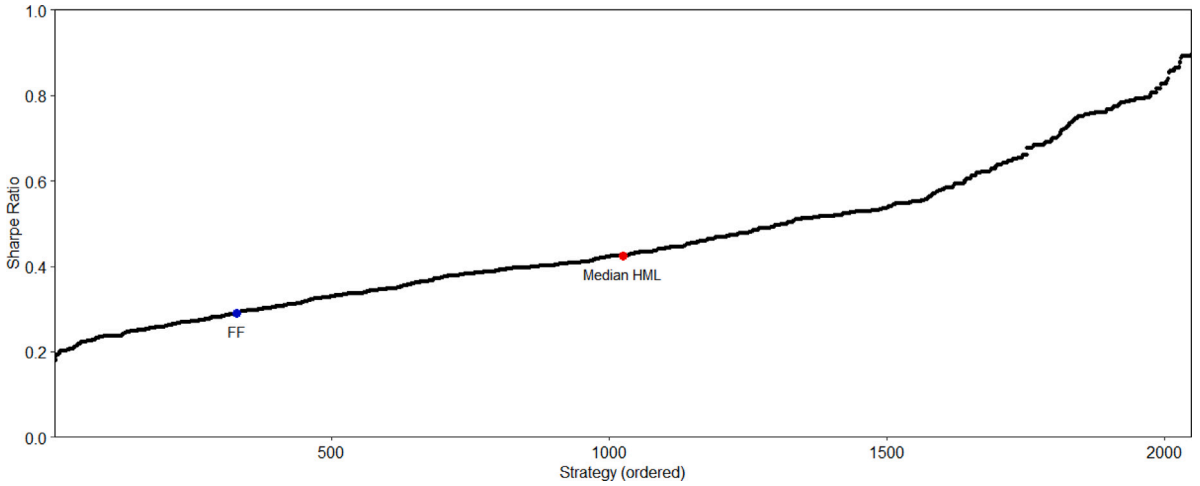

Fig. 1. Construction choices and Sharpe ratios of the HML factor. This figure plots annualized gross Sharpe ratios (y-axis) for long–short factor returns, where a factor is constructed by using 2048 different factor construction methods. The x-axis shows the 2048 different versions of the value factor ordered from low Sharpe ratios to high. The upper dot (Median HML) shows the median Sharpe ratio for the HML factor in our sample. The lower dot (FF) shows the Sharpe ratio using the construction choices mentioned in the original study. The sample runs from July 1972 until December 2021.

including or excluding microcaps, (9) independent or dependent sorts, (10) industry neutralization or not, and (11) including or excluding utility firms.

To study how important these choices are, and which choices matter most, we construct several factor models using each possible combination of choices, which leads to 2048 (2^{11}) construction combinations. For our analysis, we focus on the factor models of Fama and French (2015, 2008) augmented with momentum, the Q-factor model of Hou et al. (2015), the six-factor model of Barillas and Shanken (2018), and the factor model of Daniel et al. (2020a). Our analysis centers on maximum Sharpe ratios as these allow us to assess both individual factors and factor models (Barillas and Shanken, 2017, Fama and French, 2018). Based on U.S. stock returns, we find that factors exhibit large variation in Sharpe ratios within our set of possible construction methods. To illustrate, Fig. 1 shows the gross annualized Sharpe ratio of the canonical value factor (HML) of Fama and French (1993) for the 2048 possible construction choices. The median Sharpe ratio across the choice set is 0.43. The figure shows that the variation in obtained annualized Sharpe ratios is substantial. Depending on how we create the HML factor, Sharpe ratios vary between 0.18 and 0.89. Our paper shows that the same design choices can also strongly affect the Sharpe ratios of the other factors that we consider, which are factors relating to size, profitability, investments, momentum, return on equity, financing, and the post-earnings announcement drift.

The NSEs in our setting can be defined as the standard deviation of the generated Sharpe ratios across the possible construction methods. We find that these NSEs are sizable relative to standard errors across all factors. In multiple cases, the NSEs exceed the standard errors. For example, the NSE for the post-earnings announcement drift factor is 0.10, whereas the standard error ranges between 0.03 and 0.11. The average ratio of the NSE to the standard error across factors is 1.06. As such, factor returns are not only a function of their sorting characteristic, but also a function of their construction choices.

Non-standard errors can also materially impact factor model performance and comparison across models. Traditionally, factor models have been compared against each other based on factor construction methodology choices that differ amongst factors. For example, Ahmed et al. (2019) compare several factor models without keeping research design choices constant across factor models. Our paper performs model comparison exercises given a specific set of construction choices. We repeat this exercise for each of the 2048 combinations of choices. As such, variation in factor model performance due to variation across choices is kept constant in our model comparison exercise. We use the maximum (squared) Sharpe ratio as the main metric when comparing factor models. Barillas and Shanken (2017) show that for models with traded factors, the extent to which each model is able to price factors in the other model is what matters for model comparison, not the test assets. They propose the use of maximum squared Sharpe ratios as a model comparison metric, which Fama and French (2018) use to evaluate their 3-factor, 5-factor, and 6-factor models.[1] In 54.0% of the combination sets, we find that the Barillas and Shanken (2018) factor model emerges as the dominant factor model, followed by the factor model of Fama and French (2018) (28.6%). Using out-of-sample simulations, following Fama and French (2018), we find that the factor model of Barillas and Shanken (2018) remains the dominant factor model in 41.7% of the combination sets.

We further find a large discrepancy in optimal mean–variance weights within factor models. Moreover, our findings indicate that economic significance, i.e., how much gain could be realized by a mean–variance investor, is sensitive to construction methods. In additional tests, we study whether variation in construction choices affects factor exposure, liquidity, and transaction costs of

[1] Barillas et al. (2020) compare a range of models using the maximum squared Sharpe ratio and find that a variant of the Fama and French (2018) 6-factor model, with a monthly updated version of the value factor, emerges as the dominant model.

a portfolio. We again find that portfolio construction methods matter. For example, including microcaps leads to portfolios with higher illiquidity than excluding microcaps, which consequently results in higher risk-adjusted gross returns.

Our paper further highlights that there is large variation in the extent to which a particular choice matters. Particularly important choices are those concerning the use of NYSE or NAN breakpoints, including or excluding micro stocks, using industry-adjusted characteristics or not, and the rebalancing frequency. A better understanding of the design choices that matter most allows researchers to more effectively show the robustness of their findings. More specifically, one potential way forward is to consider these choices in a "specification check" (Brodeur et al., 2020, Mitton, 2022), in which the distribution of the results from the combinations of this limited set of possibilities are reported.

Another potential way forward is for studies to be more uniform in their choices. There is a tradeoff involved in this aspect. Variation in design choices allows researchers to customize samples, or to include the whole market, to test specific research questions. Moreover, variation reduces the chance of missed discoveries. On the other hand, though, too much variation severely complicates the comparison of results across papers. We find that a more uniform design (by just keeping a few of the choices constant) can substantially reduce non-standard errors in portfolio sorts, greatly facilitating the interpretability of the presented results. In other words, a main takeaway of our paper is the differential importance of design choices. Multiple design choices do not have large effects, and for these choices researchers could be granted freedom in customization. For the design choices that do have considerable effects, uniformity provides advantages and the asset pricing field would benefit if researchers clearly describe these choices, including motivations when deviating from the norm.

Our paper relates to empirical studies on the replicability of market anomalies. Earlier work by Hou et al. (2019) shows that the performance of factors is sensitive to the breakpoints being used, and Hou et al. (2020) show that earlier results are affected by the inclusion of microcaps.[2] Our analysis centers on a variety of potential drivers of the variation in outcomes. Our setup links to other recent papers studying design choices in asset pricing, particularly on the value premium. Kessler et al. (2020) take a practitioner's perspective to show the impact of design choices on the value premium on the S&P 500, while Hasler (2024) shows that alternative choices lead to a value premium that is smaller than originally thought. More generally, Hasler (2023) concludes that statistical biases from research decisions can explain around a fifth of the return predictability in the literature. Our goal is to get an idea about the magnitude of non-standard errors by assessing the importance of a range of construction choices, and we additionally aim to compare factor models on an "apples-to-apples" basis. Walter et al. (2023) also study non-standard errors in asset pricing. Like us, they conclude that non-standard errors are substantial, but they do not examine factor models, like we do in this study.

The remainder of this paper is organized as follows. We describe the data and factor models in Section 2. Section 3 describes the empirical variation in sorting methods of papers on factor models. Section 4 outlines the other design choices. In Section 5, we examine the importance of factor construction choices for Sharpe ratios and calculate non-standard errors. Section 6 examines whether factor construction choices impact model selection exercises. Section 7 shows how different construction methods affects several key portfolio characteristics. Section 8 concludes.

2. Constructing factor models

We obtain monthly returns and prices for U.S. equities from the Center for Research in Security Prices (CRSP). Accounting information is retrieved from the Compustat Annual and Quarterly Fundamental Files. Our sample consists of stocks listed on the NYSE, AMEX, and Nasdaq with share codes 10 or 11, which limits our sample to common stocks. The sample period spans January 1972 to December 2021.[3] We construct multiple factor models, originating from Fama and French (2015), Hou et al. (2015), Fama and French (2018), Barillas and Shanken (2018), and Daniel et al. (2020a). Table 1 summarizes the factors of each factor model and their key construction choices as used in their original studies.[4] The market factor is a part of all models. The Fama and French (2015) 5-factor model (FF5) consists of the market, size (SMB), value (HML), profitability (RMW) and investment (CMA) factors. The factors are constructed by using a 2×3 independent sort between size and the characteristic. The size sort uses a median breakpoint, and the sorting characteristic is split by the 30th and 70th percentile, all on the NYSE universe. All FF5 factors are rebalanced yearly. The FF6 model augments the 5-factor model by adding the momentum (UMD) factor. The UMD factor differs only in the rebalancing, which is monthly. In addition, we construct a cash-based version of the RMW factor (named $RMW(CP)$) for both models, following Fama and French (2018). This results in models that we abbreviate as $FF5_c$ and $FF6_c$. The Q factor model of Hou et al. (2015) consists of the market, size, investment (IA), and return on equity (ROE) factor. These factors are derived from a 2x3x3 independent sort. Barillas and Shanken (2018) combine factors from the FF models and Q model into a six-factor model (BS), consisting of the market factor, size factor, a monthly-updated value factor, the momentum factor, the growth in book assets factor, and the return on equity factor. Daniel et al. (2020a) (DHS) construct a three-factor model consisting of the market factor, financing

[2] Other related work includes McLean and Pontiff (2016), who test anomalies out-of-sample and find that the performance of identified anomalies diminishes after publication. Harvey et al. (2016) derive threshold levels to take into account potential data mining. Based on a multiple testing framework, they find that many anomalies are likely false discoveries. Linnainmaa and Roberts (2018) find that a similar conclusion can be drawn when examining pre-sample periods. Hou et al. (2020) test 452 anomalies by using a single factor construction procedure. They find that around two-thirds of the anomalies fail to replicate, even if they do not adjust for multiple hypothesis testing. On the other hand, Yan and Zheng (2017) use a bootstrap approach to evaluate fundamental-based anomalies and find that many fundamental signals are significant predictors of cross-sectional stock returns, even after accounting for data mining. Chen and Zimmermann (2022) and Jensen et al. (2023) show that they are able to successfully reproduce the majority of asset pricing factors.

[3] The starting year is 1972 as we require quarterly earnings announcements dates (to construct the price earnings announcement drift factor) and quarterly book equity data (to construct the return on equity factor).

[4] Definitions of the sorting variables are provided in the Appendix.

(FIN) factor, and the post-earnings announcement drift (PEAD) factor. Both the FIN and PEAD factor use 20–80 breakpoints in the characteristic dimension.[5] The PEAD factor is rebalanced monthly.

3. Variation in sorting methods in factor models

Researchers face a large number of methodological decisions when testing hypotheses. To examine the methodological choices that have been made in the empirical asset pricing literature focusing on portfolio sorting, we start by discussing choices made in studies on factor models.

3.1. Characteristic breakpoints

Common practice in academic finance literature has been to create portfolios by sorting on characteristics associated with expected returns. Various breakpoints have been proposed to create long–short portfolios. One standard procedure is to construct factors using a 2×3 sorting procedure as in Fama and French (1993). First, stocks are sorted by their market capitalization, whereby stocks are split into "small" and "big" classifications based on the NYSE median break-point. Second, and independently, stocks are sorted on their characteristic, whereby stocks are classified into "high" and "low" based on the 30th and 70th percentile (calculated over the NYSE universe) of the characteristic. The intersection of these classifications results into six portfolios, from which the spread portfolio is derived.

The 30th and 70th percentile breakpoints are thus one popular choice, used in, for example, the Fama–French models (Fama and French, 2018) and in the Q factor model (Hou et al., 2015). However, many others have chosen to deploy the 20th and 80th percentile to sort portfolios in the characteristic dimension. For example, Stambaugh and Yuan (2017) create mispricing factors by using the 20th and 80th percentile as characteristic breakpoints, without clear motivation for their breakpoints. Likewise, Daniel et al. (2020a) construct behavioral factors using the 20–80 breakpoints for some factors (like the PEAD factor), while using 30–70 breakpoints for the net share issuance factor. A choice for 20–80 breakpoints choice results in portfolios with more extreme characteristics. If expected returns linearly increase in those stock characteristics, such portfolios will tend to have higher average returns. Hence, factors that are constructed by using different breakpoints should not be compared directly. We construct different versions of factors where we either use the 30th–70th breakpoint or the 20th–80th breakpoint in the characteristic dimension.[6]

3.2. Breakpoints universe

Another common practice when forming portfolios is the use of NYSE-AMEX-Nasdaq (NAN) breakpoints, instead of NYSE-based breakpoints. Stocks listed on the NYSE typically have a larger market capitalization compared to those on the AMEX and Nasdaq. Furthermore, microcaps have the largest cross-sectional standard deviations of returns and stock characteristics (Fama and French, 2008). When portfolios are constructed using NAN breakpoints, microcaps may be overrepresented in extreme deciles. Consequently, this biases average spread returns upwards. However, by using NYSE breakpoints, the extreme deciles are more balanced in terms of market capitalization. Recently, Hou et al. (2020) advocate the use of NYSE-based breakpoints in portfolio sorts as the benchmark choice. Nevertheless, in the broader empirical asset pricing literature, the use of NAN breakpoints remains popular. In addition, multiple studies construct spread portfolios using NAN breakpoints, and subsequently estimate risk-adjusted returns based on asset pricing models containing factors constructed using NYSE breakpoints. This does not pose a fair comparison, and may lead to an overestimation of risk-adjusted returns. Amongst factor models, the factor models of Fama and French (1993, 2015, 2018), Hou et al. (2015), and Daniel et al. (2020a) use NYSE breakpoints in their portfolio sorting. On the other hand, Stambaugh and Yuan (2017) construct mispricing factors using NAN breakpoints.[7] In our own analysis, we construct factors both using NAN-breakpoints as well as NYSE-breakpoints.

3.3. Double and triple-sorts

Many factor models include factors constructed using 2×3 portfolio sorting, inspired by Fama and French (1993). On the other hand, the investments (IA) and profitability (ROE) factors from Hou et al. (2015) are constructed using the $2 \times 3 \times 3$ independent sorting procedure, as they independently sort on market capitalization, the annual change in total assets, and the quarterly return on equity. One disadvantage with a $2 \times 3 \times 3$ sort is that it is not always clear which additional factor to sort on in the third dimension. In case of the Q factor model, there are theoretical arguments why the ROE and IA factors should be orthogonalized: the negative relation between investment and cost of capital is conditional on return on equity. In addition, the positive relation between return on equity and cost of capital is conditional on the level of investment. Hence, Hou et al. (2015) have an economic rationale to use the $2 \times 3 \times 3$ sorting methodology. However, there is no theoretical guidance on how to construct FF-factors or DHS-factors using a $2 \times 3 \times 3$ sort, or guidance on which additional characteristic should be added in the $2 \times 3 \times 3$ sort. This additional dimension leads to another degree of freedom, where the researcher has a wide range of options to select from. Another disadvantage of the $2 \times 3 \times 3$ independent sorting methodology is that it may result in sparse or even empty portfolios. This is less likely to occur with independent 2×3 sorts.[8] Ideally, motivations for $2 \times 3 \times 3$ sorting are given and factors should be compared on an equal basis.

[5] The FIN factor is a combination of composite equity issuance (CSI) and net equity issuance (NSI). The former uses a 20–80 breakpoint.

[6] Rather than using double-sorted factors to control for size effects, most studies construct univariate sorts using decile breakpoints. Such factor portfolios will tend to have even higher average returns and non-standard errors would only increase if we would also include decile sorts in our broader analyses.

[7] In robustness tests, the authors do impose NYSE breakpoints in combination with a five dollar price filter.

[8] Sparsity can be avoided by using dependent sorts, as we discuss in Section 4.3.

For example, Ahmed et al. (2019) and Detzel et al. (2023) conduct model comparison tests, comparing the 2 × 3 × 3-based Q-factor model to the 2 × 3-based Fama–French factor model. We construct factors considering both a 2 × 3 sort and a 2 × 3 × 3 sort.[9]

3.4. Rebalancing frequency

Another potentially relevant choice for factors and factor models is when to rebalance factors, i.e., the rebalancing frequency. This typically depends on how frequent stock-level information is being updated. For example, factors extracted from the annual files of Compustat are typically updated once a year. On the other hand, price-related information is available at the monthly, daily, and even intraday frequency. One advantage of using a higher rebalancing frequency is that the researcher is able to exploit timely information. This is especially relevant for factors with short-term price predictability. On the other hand, increasing rebalancing frequency can increase portfolio turnover and thereby also transaction costs. Detzel et al. (2023) argue that prior studies on factor model selection have a tendency to pick factors that have higher rebalancing frequency, which accounts for the benefits of frequent updating while ignoring the costs involved.

Turning to factor models, Fama and French (1992) and their subsequent models update their portfolios once a year using information from end of June of the current year t. The Q-factor model of Hou et al. (2015) rebalances its factors monthly since ROE changes each month after earnings releases. Likewise, Daniel et al. (2020a) construct the PEAD factor on a monthly basis. The value factor used in Barillas and Shanken (2018) is also updated monthly. In our analyses, we consider both the use of monthly and annually rebalanced factors.

3.5. Financial firms

One of the most pervasive filters used in empirical asset pricing occurs in the industry composition of factor returns, typically focused on the exclusion of firms in the financial sector (standard industrial classification (SIC) codes between 6000 and 6999). A common argument is that financial services firms are fundamentally different in that these firms are more regulated and tend to have higher leverage and increased sensitivity to financial risks. Fama and French (1992) explicitly exclude financial firms due to their leverage since this may not indicate the distress typically associated with high leverage for non-financial firms. Firms in the financial sector constitute a significant portion of the CRSP universe. For example, in our sample, we find that 18.9% of the stocks belong to the financial services sector. As such, excluding financial services firms implies that a considerable part of the investment universe is excluded, which in turn affects factor returns. Foerster and Sapp (2005) find that excluding financial service firms from empirical asset pricing tests can impact the corresponding inferences. It may influence both the identification of the number of risk factors found to be significant and the corresponding betas. Alternatively, if a researcher is worried about unintended exposure towards financial firms when constructing factors, imposing sector neutrality in portfolio sorts can also alleviate such concerns without discarding a part of the investment universe. Most of the literature, including Daniel et al. (2020a), follows Fama and French (1992) when it comes to excluding financial firms. Still, many other studies do include financial firms in their samples. For example, Stambaugh and Yuan (2017) include financial firms in their sample, without explicitly stating why. In our own analysis, we construct factors using investment universes that either include or exclude financial services firms.

3.6. Negative book equity value

The book equity of a firm represents the difference between a firm's assets and liabilities, and measures the equity held by a firm's shareholders. Given the limited liability structure of a firm, shareholder value cannot seem to be negative. Nevertheless, in practice, firms can have negative book equity, for example if firms have large negative earnings or goodwill impairments. Fama and French (1993) exclude negative BE stocks arguing that such occurrence is rather rare before 1980. However, negative BE firms became more prevalent after the 1980s. For example, in our sample, the frequency of negative book equity firms increases from 1.1% in July 1972 to 4.6% in December 2021. As such, their inclusion or exclusion may yield a different stock universe, and consequently yield different results. For example, Brown et al. (2008) show that negative book equity firms are disproportionately represented in extreme growth and value sectors. There is no consensus in the empirical asset pricing literature whether to include or exclude negative book equity firms. Many papers tend to follow Fama and French (1993) in this regard, but others, such as the mispricing factors of Stambaugh and Yuan (2017), are constructed using a stock universe that includes negative book equity firms. In our own analysis, we construct factors using stock universes that either include or exclude negative book equity firms.

4. Other construction choices

This section outlines other portfolio sorting decisions that asset pricing researchers frequently make in their studies. These choices are relevant because the portfolios resulting from these choices are oftentimes evaluated against factor models. Also, although these choices are not directly related to the range of factor models that we consider in this paper, they might be used in the development of future factor models.

[9] For simplicity, when we construct FF or DHS factors via a 2 × 3 × 3 sort, we use the book-to-market ratio as the additional sorting characteristic in the third dimension.

We do not consider the sample period as a construction choice, as the convention is to start in the year when all relevant data become available and finish in the most recent year with full data availability (at the time of the analysis). The download date of the data can matter: Akey et al. (2022) show that Fama–French factor returns vary across different factor vintages. In addition, we do not consider the weighting choice, such as value-weighting versus equal-weighting. The main reason is that it is already common practice to show robustness for the alternative weighting scheme. In our own analysis, we construct factors using value-weighting (based on market capitalization). No factor model in our paper uses equal-weighted factors in their original set-up. Value-weighted portfolios typically serve as a benchmark against which portfolio managers are evaluated, which highlights the relevance of value-weighting in practice.

4.1. Stock price filters

Stock price filters are a common choice that researchers face within the empirical asset pricing literature. A researcher can choose to exclude stocks based on any price filter, or to not exclude stocks based on prices. In the former case, a minimum price has to be decided to exclude "penny" stocks. Typically this price is set to 1 or 5 dollar. A few studies also exclude highly priced stocks, such as stocks with a price above 1000 dollars (as in Bali et al., 2017). The aim of excluding penny stocks is to primarily avoid micro-structure effects. On the other hand, a stock with a low price is not per se a micro-cap, and can still be considered as liquid. For example, in our sample of stocks, 60% of stocks per month have a market capitalization below the 20th NYSE size percentile and 42% of those stocks have a price below 5 dollar, on average. As such, imposing only a price filter does not necessarily exclude all microcaps, and may even include liquid stocks. Nevertheless, imposing price filters in portfolio sorts remains to be widely used. More notably, multiple studies construct long–short anomaly returns while imposing price filters, and consequently compare this to factor models constructed without price filters. In our analysis, we either set no price filter or we use a 5 dollar price filter.

4.2. Microcaps

We also consider the inclusion and exclusion of microcaps as a construction choice. Microcaps are typically defined as stocks with a market capitalization below the 20th percentile for NYSE stocks. Fama and French (2008) find that microcaps account for 60% of the number of stocks, but only capture 3% of the total market capitalization. In addition, they find that microcaps have the highest cross-sectional volatility of returns and show large dispersion in sorting characteristics. From a practical perspective, these small stocks are out of reach for many (institutional) investors. In addition, microcaps are more expensive to short due to high shorting fees (Drechsler and Drechsler, 2014), they may be illiquid, and they have relatively high transaction costs (Novy-Marx and Velikov, 2016). Nevertheless, microcaps are typically included in many studies. Hou et al. (2020) find that including microcaps can have important effects on reported results. Excluding microcaps increases the median market capitalization, reduces typical return volatility, and increases the market share of stocks below the median. In our analysis, we either remove stocks below the 20th NYSE market cap percentile or not.

4.3. Independent versus dependent sorting

Independent sorting is the most commonly used sorting procedure deployed in the literature. A major drawback is that independent sorting may result in sparse portfolios, with the consequence that a factor portfolio is not well-diversified. In some cases, independent sorting may even result in empty portfolios, which is especially an issue in international or smaller samples (Ang et al., 2006, Novy-Marx, 2013, Wahal and Yavuz, 2013). Dependent sorting alleviates the problem of sparse portfolios by sequentially stratifying stocks into portfolios. However, implementing a dependent sorting procedure raises the question of what order of the sort should be used, especially when sorting on more than two factors. For the 2×3 procedure, the standard is to first sort on size, and then on the sorting characteristic, i.e., there is little degree of freedom in this choice. However, when we consider a $2 \times 3 \times 3$ dependent sort, it is not clear what the ordering should be, allowing for a wider playing field.

4.4. Industry hedging

Additionally, we consider industry hedging as a construction choice. The unconditional predictive power of stock characteristics may stem from their across-industries component or from their firm-specific (within-industries) component, or from both (Ehsani et al., 2021). A consequence of unconditional sorting is that factor portfolios obtain differential exposure towards specific industries. To illustrate, constructing the unconditional value factor overweights sectors that contain stocks with high book-to-market ratios, such as utility firms in the long leg, whereas the short value leg gets excess exposure towards technology stocks.

Daniel et al. (2020b) suggest that sorting stocks, unconditionally, tends to pick-up unintended (industry) risks, generating portfolios that are no longer mean–variance efficient. Sector-concentrated portfolios are more volatile because stocks within the same sector are highly correlated. Under-diversification due to these exposures do not implicitly reveal information about the expected returns of factors and hedging these exposures is a choice that can be made in order to improve risk-adjusted returns.[10] A comparison of the standard and industry-hedged factors shows that industry adjustment often improves factor performance (Novy-Marx, 2013).

[10] Especially practitioners typically add industry constraints in portfolio construction processes to avoid concentration risks.

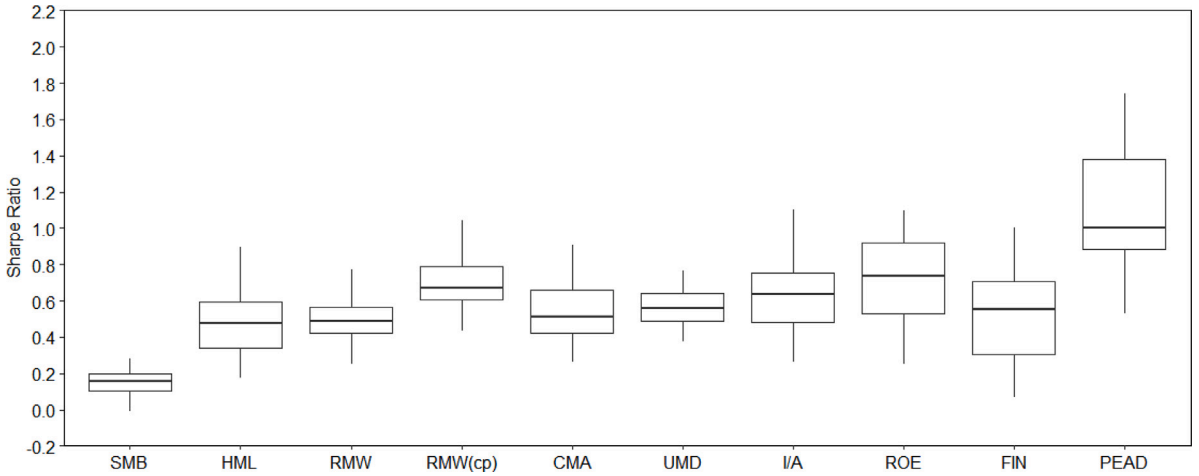

Fig. 2. Sharpe ratio variation within factors. This figure plots the distribution of annualized Sharpe ratios for long–short factor returns, where a factor is constructed 2048 times by using the 2048 different factor construction methods. The black solid line within the box plot shows the median Sharpe ratio. The upper (lower) bound shows the 75th (25th) percentile. The factors and their definitions are from Table 1. The sample runs from July 1972 until December 2021.

We construct industry-hedged factors, in addition to unhedged factors, by normalizing the sorting characteristic into an industry-adjusted characteristic as follows:

$$S^{*}_{i,t} = (S_{i,t} - \bar{S}_{i,j,t})/(S_{max,j,t} - S_{min,j,t}) \quad (1)$$

$S_{i,t}$ ($S^{*}_{i,t}$) denotes the (industry-adjusted) sorting characteristic. $\bar{S}_{i,j,t}$, $S_{max,j,t}$ and $S_{min,j,t}$ are equal to the cross-sectional mean, maximum and minimum, respectively, of the sorting characteristic S for industry j. We use the Fama–French 12-industry classification.

4.5. Utility firms

Lastly, we consider the inclusion or exclusion of utility firms in empirical asset pricing studies, which are firms with SIC codes between 4900 and 4999. Utility firms typically engage in the generation, transmission and/or distribution of electricity, gas, or steam, while the category also includes firms active in waste management. On average, 3% of the firms in our sample are classified as utility firms. The reason for excluding public utility firms may be due to their linkage to the government. Public firms often are not profit-orientated and are highly affected by governmental decisions and regulations. Their economic role is to serve public tasks, so their business model also differs from other private companies. Non-public utility firms empirically also have a very high leverage and therefore untypical high book-to-market ratio, which results in a high sensitivity to interest rate changes. As such, utility firms typically tend to be considered as value-like stocks. Excluding utility stocks may tilt the universe, as a whole, towards growth stocks. Amongst well-known factor models, none of them exclude utility firms. Still, there are studies focusing on an anomaly return that exclude utility firms from their sample. Again, differences in choices across factors do not allow for a fair comparison. To illustrate, Hirshleifer et al. (2018) exclude utility firms, construct a factor sorted on "innovative originality", and compute risk-adjusted returns relative to factor models constructed on universes that include utility firms.

5. The impact of construction choices and the size of non-standard errors

In this section, we examine the impact of portfolio design choices on Sharpe ratios. In addition, in this section we compute non-standard errors and compare these with estimated standard errors. This section concludes with an analysis of potential reductions in non-standard errors.

5.1. Construction choices and Sharpe ratios

We construct each factor in our sample 2048 times by using the 2048 different factor construction methods. Fig. 2 shows the Sharpe ratio distribution across sets of construction choices for each factor, based on long–short factor returns. Across and within factors, we observe substantial variation in Sharpe ratios. For example, the Sharpe ratio of the CMA factor ranges between 0.26 and 0.91, the Sharpe ratio of the UMD factor ranges between 0.38 and 0.77, and the Sharpe ratio of the ROE factor ranges between 0.25 and 1.10. Hence, in relative terms, the Sharpe ratio can more than double, or even quadruple, depending on design choices, and this applies to the far majority of factors. In absolute terms, the PEAD factor shows the largest variation in absolute terms, ranging from a Sharpe ratio of 0.53 to 1.74. Overall, these results imply that construction choices matter.

Table 1
Factor models. This table lists the non-market factors used by asset pricing models, indicated by a ✓. It also lists properties of the factor construction methodology: the sorting characteristic, the breakpoints (BP), the rebalancing frequency (Rebal.), and the sorting method (Sort.). FF5 (FF6) denote the Fama and French (2015) five-factor model (augmented with UMD). $FF5_c$ and $FF6_c$ denote versions of the FF5 and FF5M, respectively, that use cash-based operating profitability instead of accruals operating profitability, based on Fama and French (2018). Q4 denotes the Hou et al. (2015) four-factor model. BS6 denotes the Barillas and Shanken (2018) six-factor model. DHS3 denotes the Daniel et al. (2020a) three-factor model.

Factor	Sorting characteristic	BP	Rebal.	Sort.	Factor models						
					FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS3
SMB	Market capitalization	50–50	Annual	2 × 3	✓	✓	✓	✓	✓	✓	
HML	Book-to-market	70–30	Annual	2 × 3	✓	✓	✓	✓		✓	
RMW	Accruals operating profitability	70–30	Annual	2 × 3	✓	✓					
RMW_c	Cash operating profitability	70–30	Annual	2 × 3			✓	✓			
CMA	Growth in book assets	70–30	Annual	2 × 3	✓	✓	✓	✓			
UMD	$R_{t-12,t-2}$	70–30	Monthly	2 × 3		✓		✓		✓	
I/A	Growth in book assets	70–30	Monthly	2 × 3 × 3					✓	✓	
ROE	Quarterly returns-on-equity	70–30	Monthly	2 × 3 × 3					✓	✓	
FIN	Net and composite share issuance	80–20	Annual	2 × 3							✓
PEAD	4-day CAR earnings announcements	80–20	Monthly	2 × 3							✓

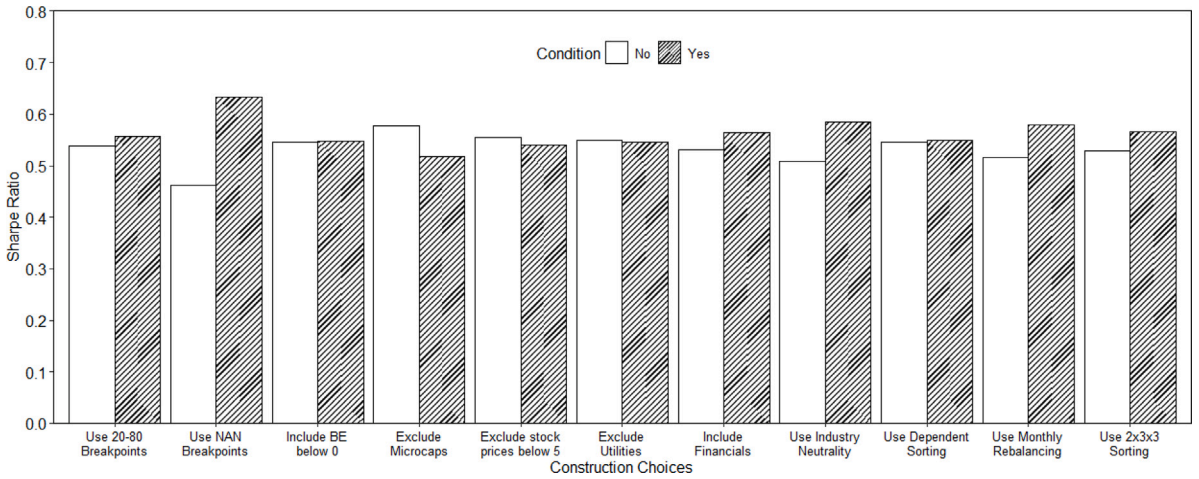

Fig. 3. Construction choices and Sharpe ratios. This figure shows the impact of construction choices on the annualized Sharpe ratio averaged over factors. "Use 20–80 Breakpoints" indicates the use of the 20th (30th) and 80th (70th) percentile as breakpoints if the condition is "Yes" ("No"). "Use NAN Breakpoints" uses the NYSE-AMEX-Nasdaq (NYSE) cross-section to make breakpoints if the condition is "Yes" ("No"). "Include BE below 0" indicates whether stocks with negative book equity are included ("Yes") or not ("No"). "Exclude Microcaps" indicates whether we exclude (include) stocks with the smallest 20% market cap if the condition is "Yes" ("No"). "Exclude stock prices below 5" indicates whether we exclude stocks with prices below 5 dollar (at the end of the previous month) if the condition is "Yes" or not ("No"). "Exclude Utilities" means that companies in the utility sector are excluded ("Yes") or not ("No"). "Include Financials" means that financial companies are included ("Yes") or excluded ("No"). "Use Industry Neutrality" means that portfolio sorts are constructed using industry-adjusted characteristics ("Yes") or not ("No"). "Use Dependent Sorting" refers to the use of dependent sorting ("Yes") or independent sorting ("No"). "Use monthly rebalancing" indicates that we use the one-month lagged characteristic ("Yes") or characteristic data from last June ("No"). "Use 2 × 3 × 3 Sorting" equals "Yes" ("No") if 2 × 3 × 3 (2 × 3) sorts are used. Monthly factor returns are constructed using data from July 1972 to December 2021.

Next, we examine how specific construction choices, in isolation, affect maximum Sharpe ratio estimates. Fig. 3 shows annualized maximum Sharpe ratios by construction choice, averaged over factors. The white bar corresponds to the construction choices made in Fama and French (1993) and serves as a benchmark. The dashed bar corresponds to the choice stated on the x-axis.

We first vary the breakpoints that are used to classify high and low characteristics. The first two bars use the 30th–70th percentile (white bar) or the 20th–80th percentile (dashed bar). The former case yields an average annualized Sharpe ratio of 0.54, whereas 20–80 breakpoints yield a Sharpe ratio of 0.56. Intuitively, if expected returns are monotonically related to a given stock characteristic, then taking positions in stocks with more extreme characteristics would naturally result into higher returns and Sharpe ratios. Using NAN breakpoints instead of NYSE breakpoints improves Sharpe ratios from 0.46 to 0.63, which is the largest increase within our set of choices. This choice thus comes out as important, where NYSE breakpoints represent the conservative choice. Another important choice is the choice whether to exclude microcaps or not. Excluding microcaps decreases the average Sharpe ratio from 0.58 to 0.52, which makes excluding these firms the conservative choice. It can further be seen that eliminating industry exposures from factor returns substantially increases Sharpe ratios, which is in line with Daniel et al. (2020b). In addition, using monthly rebalancing to construct factors increases the Sharpe ratio from 0.52 to 0.58 relative to using yearly rebalancing. Finally, using a 2 × 3 × 3 sort, on average, improves Sharpe ratios relative to the 2 × 3 sort. Choices that do not lead to substantially different average Sharpe ratios include choices related to negative book equity firms, a five dollar price filter, and utility firms. Including financial firms increases

the Sharpe ratio, on average, from 0.53 to 0.56. The Sharpe ratios for independent and dependent sorts are approximately similar. Overall, our findings imply that some construction choices can materially affect factor performance, especially those concerning NYSE breakpoints, micro stocks, industry-adjusted characteristics, and rebalancing frequency, while other choices do not have much of an impact.

5.2. Non-standard errors versus standard errors

Based on the above analyses, non-standard errors seem to be sizable. Traditionally, the focus of the empirical finance literature has been on standard errors, resulting from a data-generating process drawing samples from a population. That is, sampling uncertainty leads to standard errors when estimating population parameters, such as the mean and volatility of returns. Non-standard errors result from an evidence-generating process, which translates the sample into evidence, and which adds an additional layer of error (Menkveld et al., 2023).

We model non-standard errors as the cross-sectional standard deviation across hypothetical researchers who all use different sets of construction choices. We thus obtain one non-standard error per factor, equal to the standard deviation of the 2048 different Sharpe ratios for that factor. To compare non-standard errors with standard errors, we estimate standard errors by block-bootstrapping each factor's return for a given set of construction choices. The standard error is the standard deviation of the Sharpe ratio obtained from block-bootstrapping a factor, and we block-bootstrap each series 10.000 times. Subsequently, we average the standard errors for each factor across all choices. We show the results in Fig. 4. The white bars indicate the NSE for each factor and the dashed bars denote the estimated standard errors. Besides the average standard errors, we also plot the minimum and maximum standard errors.

We find that non-standard errors are sizable relative to standard errors, across all factors. In 4 out of 10 factors, we find that the non-standard error is considerably larger than the average standard error. These factors are IA, ROE, FIN and PEAD. The non-standard errors are relatively low for SMB. The non-standard error is highest for the PEAD factor (i.e., 0.10, whereas the standard error ranges between 0.03 and 0.11). In terms of proportions (non-standard error divided by average standard error), we find that this proportion ranges between 43% (SMB) and 197% (PEAD), with the average being 106%. Overall, we conclude that NSEs are sizable in comparison with standard errors. The average NSE to standard error ratio of 106% is also relevant in comparison to the ratio of 160% found by Menkveld et al. (2023), based on a relatively high degree of researcher discretion in their experiment with index futures trading data.

5.3. Reducing non-standard errors

Variation in design choices allows researchers to customize samples and empirical tests to tackle specific research questions. For example, researchers might be particularly interested in patterns within financial firms, or within a particular other type of firm. Allowing some variation could thus be optimal, also to reduce the chance of missed discoveries. However, high non-standard errors severely complicate the interpretability of the results by the average reader, while also potentially inducing excessive reporting of statistically significant results.

In this section, we take this trade-off into account and examine whether a limited set of restrictions could substantially reduce non-standard errors. This analysis follows from Fig. 3, which provides insights into which of the eleven design choices appear most relevant for non-standard errors. We construct two sets of potential restrictions and compare the resultant non-standard errors with those of the setting when all choices are free.

Let "Set 1" be the base case where researchers can make all of the eleven choices identified in Sections 3 and 4. In "Set 2" we exclude three choices that appear particularly important in Fig. 3: NAN breakpoints, including microcaps, and monthly rebalancing.[11] Excluding these three choices could substantially reduce uncertainty in interpreting reported results. In addition, the choices can be relatively easily justified based on economic arguments. Although approximately 60% of the stocks in the CRSP sample can be considered as microcaps, they only represent about 3% of the total market capitalization of the CRSP universe. Transaction costs for microcaps are high and liquidity is low, which makes this a segment of the market difficult for investors to invest in. Using NAN breakpoints favors micro- and small caps, leading to similarly inflated anomaly profits. Furthermore, monthly rebalancing can result into high transaction costs. Excluding these three choices leaves researchers with eight remaining design choices, or 256 possible combinations.

In "Set 3", we additionally restrict four choices that are motivated by Fama and French (1992) and Fama and French (1993): we use 30–70 breakpoints rather than 20–80 breakpoints, we exclude firms with negative book values, we exclude financial firms, and we use market equity observed in June. Not selecting 20–80 breakpoints could be defended as such breakpoints reduce portfolio breadth and could tilt towards stocks with more exposure towards a certain factor, potentially biasing the portfolio returns upward. Firms with negative book equity value might have particularly high default risk, and the relation between default risk and leverage is different for financial firms than for other firms (Fama and French, 1992). Set 3 thus only leaves four choices open: whether to impose price filters (which seems less important now that microcaps are excluded), whether to include utilities, whether to impose industry neutrality, and using dependent or independent sorts. These four choices allow 16 combinations.

[11] A fourth choice that seems important for non-standard errors is industry neutralization. The trade-off might be especially important here. Fig. 3 suggests that the conservative choice is to not use industry-adjusted characteristics. However, Daniel et al. (2020b) suggest that this tends to pick-up unintended (industry) risks, generating portfolios that are no longer mean–variance efficient. Hedging this exposure is thus a choice that might be sensible.

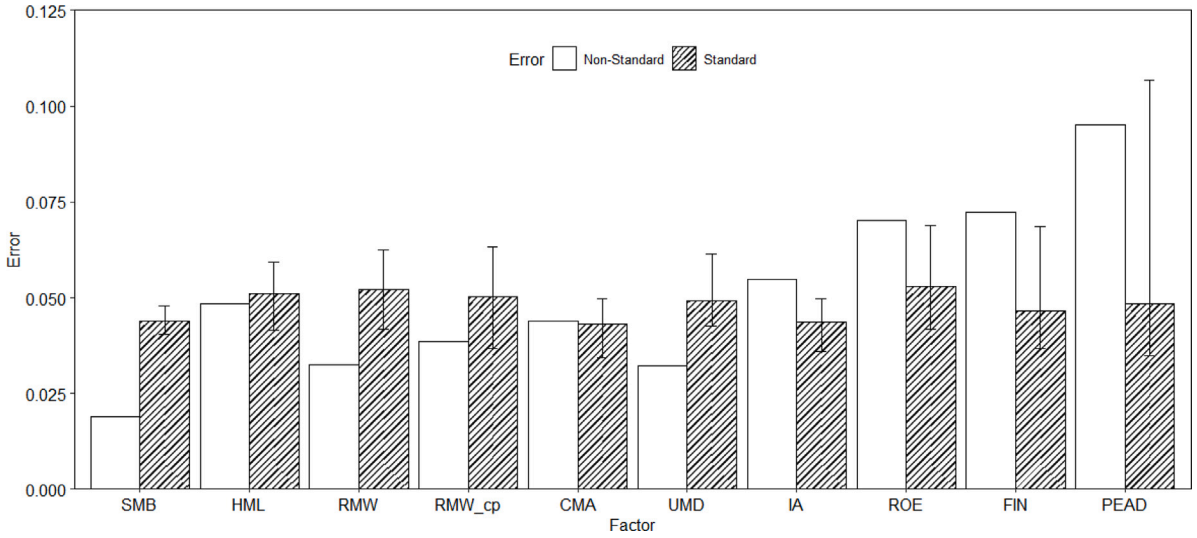

Fig. 4. Non-standard errors and standard errors. This figure plots the non-standard error (white) and standard error (dashed bar) for each factor. The non-standard error is defined as the cross-sectional standard deviation of Sharpe ratios, where the cross-section consist of all 2048 versions of a factor. The standard error is the standard deviation of the Sharpe ratio obtained from block-bootstrapping a factor, averaged over the construction choices. We block-bootstrap each series 10.000 times. The error line on the dashed bar indicates the minimum and maximum standard error within a factor. Monthly factor returns are constructed using data spanning July 1972 and December 2021.

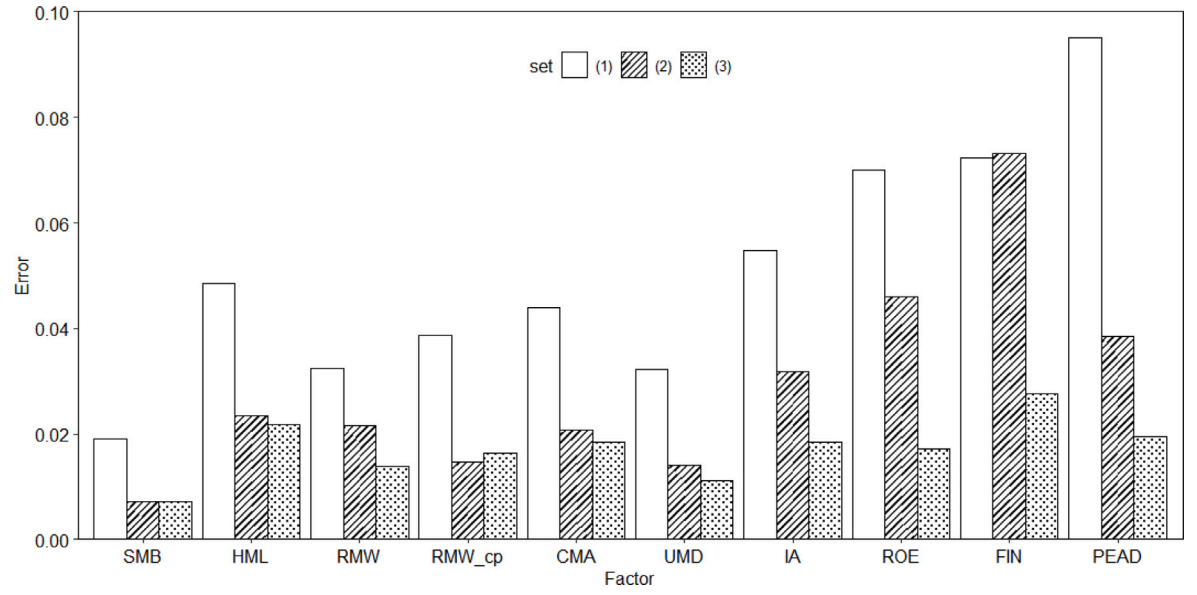

Fig. 5. Reducing non-standard errors. The figure shows the non-standard errors using three sets of research design choices. Set (1) includes all eleven construction choices. Set (2) imposes NYSE breakpoints, excludes micro-caps, and excludes monthly rebalancing. Set (3) extends on set (2) by further imposing the use of 30–70 breakpoints, the exclusion of firms with a negative book value and financial firms.

Fig. 5 shows the computed non-standard errors per factor for each set. We find that the NSE can be heavily reduced by imposing the restrictions of Set 2, i.e., the use of NYSE breakpoints, excluding microcaps, and excluding monthly rebalancing. For example, the PEAD factor has a NSE of above 0.10 using the original set of eleven choices, which decreases to about 0.04 (a 60% decline) for Set 2. On average, across factors, we find that NSEs decrease by 42% when moving from Set 1 to Set 2. On average, Set 3 leads to a 66% reduction compared to the base case. When keeping in mind that imposing restrictions hurts opportunities for customization, a relatively simple recommendation to reduce non-standard errors that follows from the above analysis is to consistently use NYSE breakpoints, exclude microcaps, and exclude monthly rebalancing. Of course, in some cases an argument can be made for not following this recommendation. For instance, researchers might have a particular interest in smaller firms, or they might want to

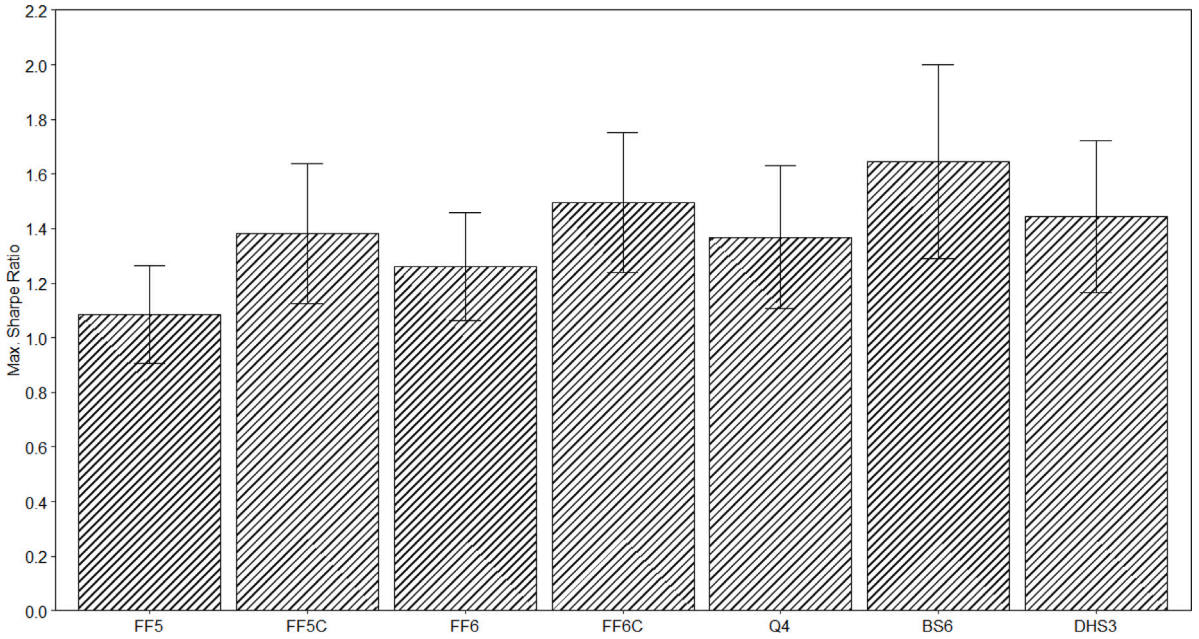

Fig. 6. Selecting factor models. This figure shows the maximum gross Sharpe ratio (annualized) from the factors from the factor models listed on the horizontal axis. The error plot shows the variation (95% confidence interval) in the maximum Sharpe ratios for a given factor model, across construction choices. The data runs from July 1972 until December 2021.

study a mechanism most applicable to illiquid stocks. Providing a clear explanation for design choices that deviate from the above recommendation in such studies appears warranted.

6. Model selection

The prior section has shown that Sharpe ratios within factors depend on a range of construction choices and that NSEs surrounding portfolio sorting can be substantial. In this section, we study the implications of NSEs for model selection exercises. In particular, we use the maximum squared Sharpe ratio as selection criteria for ranking asset pricing models. Additionally, we consider efficient frontier expansion, economic significance, and out-of-sample estimation, following Detzel et al. (2023). Prior model selection studies select factors, each based on an own set of portfolio construction choices. As these construction choices differ per factor, outcomes of model selection studies are impacted by underlying differences in the set of construction choices. In this section, we construct factors under the same set of portfolio construction choices. Thereby, we remove idiosyncratic construction effects from factors, and introduce common construction effects to evaluate them on the same basis for model selection. We then show the impact of construction choices on the outcomes of these model selection exercises.

6.1. Maximum Sharpe ratio

The ability of an asset pricing model to price assets depends on the extent to which its factors span the mean–variance efficient portfolio. When the factors of a model are mean–variance efficient, no other factor or asset can be added to improve the performance of the span of the factors. Gibbons et al. (1989) show that the gain of adding test assets to a factor model can be written as:

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

$Sh^2(f,\Omega)$ denotes the maximum squared Sharpe ratio obtained from the factors f and assets Ω, and $Sh^2(f)$ for f. α is a vector of intercepts obtained from regressing the assets Ω excess return on factor returns. Σ^{-1} is the covariance matrix of residuals from these regressions. Barillas and Shanken (2017) use the maximum squared Sharpe ratio as an indicator of model quality, since it measures how close the span of a model is to the ex-post mean–variance efficient frontier. The aim is to minimize the mispricing that an asset pricing model creates, which corresponds to minimizing the outcome of Eq. (2). Barillas and Shanken (2017) argue that $Sh^2(f,\Omega) = Sh^2(\Omega)$ when Ω consists of the entire universe of assets. In that case, minimizing the outcome of Eq. (2) corresponds to maximizing $Sh^2(f)$. Hence, model selection can be examined by comparing the maximum squared Sharpe ratio across models.[12]

[12] Detzel et al. (2023) show that when (transaction) costs are ignored, model comparison based on squared Sharpe ratios favor models with high gross performance, even when trading costs are high.

The typical approach in the literature has been to compare factors using their "original" construction method, thereby comparing factors without taking differences in construction method into account. We explicitly take into account the range of possible construction methods and compare factors on an "apples-to-apples" basis. Fig. 6 reports the average maximum Sharpe ratio of a factor model whereby we average across the possible set of construction choices. Around the average, we also plot a two standard deviation spread of the Sharpe ratio of a factor model.

The average maximum Sharpe ratio for the mean–variance optimal FF5 model is 1.09. Replacing operating profitability with cash profitability increases this value to 1.38. Adding the momentum factor further improves the average maximum Sharpe ratio to 1.50. The optimal Q4, BS6, and DHS factor models have an average maximum Sharpe ratio of 1.37, 1.65, and 1.44, respectively. Based on these averages, the preferred model would be the BS6 factor model, with the FF6C model coming second.

Differences in construction choices induce NSEs in factor premiums and subsequently in the maximum Sharpe ratio of factor models. The error bars indicate that the two-standard deviation spread in the maximum Sharpe ratio can be substantial. For example, for the BS6 model, we find a 95% confidence interval between 1.29 and 2.00. Due to the NSEs, model rankings may differ across different sets of construction choices. We find that in 54% of all choice sets, the BS6 has the largest maximum Sharpe ratio. The FF6C model has the largest maximum Sharpe ratio in 28.6% of the choice sets. These results show that even though one model can have the largest maximum Sharpe ratio in the majority of the construction choice sets, a different outcome for model selection exercises can be achieved when using other choice sets. Moreover, model rankings can especially differ when researchers make different choices across models (for example, 80–20 breakpoints for the PEAD factor but 70–30 breakpoints for factors in another model).

Table 2 reports the portfolio weights that correspond to the ex-post mean–variance efficient portfolios constructed from the candidate factor models, where we average the weights across all construction methodologies. Between brackets, we report the standard deviation of the weights, based on our set of 2048 construction methods. The standard deviation can be considered as a non-standard error in the mean–variance optimal weights due to variation in construction methodologies. Since the factors are constructed in the same way, the weights can be compared directly. We find large discrepancy in optimal weights within factor models. For example, the Fama–French 5 factor model allocates 41.4% weight towards the CMA factor, on average. However, for a researcher that randomly picks a construction choice, the weight on the CMA factor varies between 35.0% and 47.8% for a two standard deviation change. HML has a small average weight of 3.3% in the 5-factor model. Interestingly, for some construction methods the HML weight is negative (−12.5% for a two standard deviation decrease) while for others it is substantially positive (19.1% for a two standard deviation increase). Hence, in some situations, it appears that one should have a short position in HML, whereas with other construction choices a mean–variance investor should hold a long position in HML. The Q4 model aims to improve on the 5-factor model by replacing the investment factors with their ROE factor, which uses more timely information (i.e., quarterly ROE data). Compared to the Fama–French models, the Q factor seems to have more stable weights, with standard deviations between 2% and 5%. For example, the I/A factor ranges between 29.4% and 48.2%, given a two standard deviation interval. On average, the PEAD factor is important in the DHS model. The model allocates on average 59.8% to the PEAD factor, with a standard deviation of 10.8%.

6.2. Efficient frontier expansion

The results from the previous sub-section indicate that model performance and its underlying weights depend on construction methods. In this subsection, we aim to measure the extent to which additional factors of a model "M1" to those of model "M0" expand the efficient frontier. To this end, we implement the multi-factor version of the generalized alpha of Novy-Marx and Velikov (2016). More specifically, we run a regression of the excess returns of the ex-post mean–variance efficient portfolio constructed from the union of M1 and M0 on the returns of the mean–variance efficient portfolio using the factors from model M0:

$$MVP_{M1\cup M0,t} = \alpha + \beta MVP_{M0,t} + \epsilon_t \tag{3}$$

Table 3 reports the results of these spanning regressions for each pair of models, averaged over all construction methods. Typically, we find that most models expand the efficient frontier when added to other models. For example, the spanning alpha of the FF5 model augmented by other models ranges between 0.11% and 0.30% per month. We especially find that adding the BS6 factors or DHS3 factors (M1) to FF models (M0) greatly improves the efficient frontier, with alphas between 0.28% and 0.30% per month.

Across construction methods, we find large standard deviations in the estimated alphas (reported within [ ]). The Q4 model, on average, expands its efficient frontier by adding other factor models. For example, adding the FF5 model to the Q4 model expands the efficient frontier, on average, with an estimated average alpha of 0.04% per month. However, the estimated alpha has a standard deviation of 0.04%. Under some construction method, the estimated alpha may thus be considerably closer to zero. Hence, in some cases it may appear that adding one factor model to other factor models expands the efficient frontier, whereas in other cases the marginal benefit of adding a factor is small or even zero. Again, our results imply that construction methods can influence model selection exercises, as indicated by the relatively large standard deviations around the spanning alphas.

6.3. Economic significance

Next, we quantify the economic significance, in Table 4, by reporting by which percentage the maximum Sharpe ratio would increase if we would add the additional factors (M1) to the base model (M0) for each pair. This exercise relates to the gain that could be realized by a mean–variance investor. In most cases, adding one model to a base model improves the Sharpe ratio of the

Table 2
Mean–variance efficient portfolio weights. This table shows the optimal weights that a mean–variance efficient investor would allocate to factors within a factor model, averaged over our set of possible construction methodologies. Within brackets, we show the standard deviation of the optimal weights that occur within our set of possible construction methods. The table shows the weights using factor returns gross of transaction costs. The sample period is from July 1972 to December 2021.

	Mkt	SMB	HML	RMW	CMA	UMD	RMW_{cp}	IA	ROE	HML_d	FIN	PEAD
FF5	20.2	5.4	3.3	29.7	41.4							
	(4.1)	(2.8)	(7.9)	(6.0)	(6.4)							
FF5C	17.0	5.4	14.9	22.4	20.9	19.3						
	(4.0)	(2.3)	(9.3)	(6.4)	(8.7)	(5.1)						
FF5C	18.2	8.6	1.8		27.1		44.3					
	(3.5)	(3.1)	(7.4)		(5.9)		(6.9)					
FF5MC	16.3	7.8	11.8		14.2	15.6	34.3					
	(3.6)	(2.2)	(8.7)		(6.9)	(4.2)	7.0					
Q4	17.4	11.2						38.8	32.6			
	(2.9)	(2.3)						(4.7)	(4.2)			
BS6	16.1	8.7				11.1		22.1	26.2	15.8		
	(3.2)	(3.4)				(7.3)		(8.8)	(6.2)	(8.7)		
DHS	16.8										23.4	59.8
	(3.5)										(8.5)	(10.8)

Table 3
Frontier expansion. This table reports the intercepts obtained from the regression $MVE_{M1UM0,t} = \alpha + \beta MVE_{M0,t} + \epsilon_t$. M0 is the "base" model, which is augmented to model $M1UM0$ by adding the factors of $M1$ to $M0$. $MVE_{M1UM0,t}$ is the corresponding mean–variance efficient portfolio obtained from the union of factors of $M1$ and $M0$. $MVE_{M0,t}$ is the mean–variance efficient portfolio of the factors from model $M0$. The t-statistics, reported within parentheses, are heteroskedasticity robust. Within brackets, we report the cross-sectional standard deviation of alpha. The table reports the results using gross returns. The data runs from July 1972 until December 2021.

M0	$M1UM0,t$						
	FF5	$FF5_c$	FF6	$FF6_c$	Q4	BS6	DHS3
FF5	0.00	0.11	0.17	0.22	0.20	0.30	0.28
	(0.00)	(4.36)	(5.42)	(6.78)	(6.09)	(8.53)	(8.56)
	[0.00]	[0.04]	[0.07]	[0.06]	[0.12]	[0.12]	[0.10]
$FF5_c$	0.01	0.00	0.14	0.13	0.12	0.23	0.24
	(1.19)	(0.00)	(5.06)	(4.90)	(4.71)	(7.34)	(8.00)
	[0.02]	[0.00]	[0.06]	[0.05]	[0.09]	[0.10]	[0.10]
FF6	0.00	0.07	0.00	0.07	0.10	0.16	0.20
	(0.00)	(3.66)	(0.00)	(3.66)	(4.08)	(5.33)	(7.06)
	[0.00]	[0.03]	[0.00]	[0.03]	[0.08]	[0.13]	[0.08]
$FF6_c$	0.01	0.00	0.01	0.00	0.06	0.11	0.17
	(0.98)	(0.00)	(0.98)	(0.00)	(3.26)	(4.37)	(6.77)
	[0.01]	[0.00]	[0.01]	[0.00]	[0.06]	[0.10]	[0.08]
Q4	0.04	0.06	0.10	0.12	0.00	0.15	0.19
	(2.13)	(3.18)	(3.99)	(4.90)	(0.00)	(5.54)	(6.59)
	[0.04]	[0.04]	[0.08]	[0.08]	[0.00]	[0.06]	[0.07]
BS6	0.02	0.03	0.02	0.03	0.00	0.00	0.19
	(1.57)	(2.34)	(1.57)	(2.34)	(0.00)	(0.00)	(6.92)
	[0.03]	[0.03]	[0.03]	[0.03]	[0.00]	[0.00]	[0.07]
DHS	0.13	0.17	0.17	0.19	0.15	0.25	0.00
	(5.15)	(6.15)	(5.78)	(6.54)	(5.69)	(8.17)	(0.00)
	[0.06]	[0.06]	[0.06]	[0.06]	[0.07]	[0.11]	[0.00]

combined model. We find that the FF5 model can be improved between 18.7% up to 59.3%, on average, by adding one of the other factor models. Adding the BS6 model to any of the FF-models could improve the maximum Sharpe ratio by between 7.6% and 42.4%, whereas adding the DHS3 model yields gains between 25.7% and 59.3%. These results indicate that the FF models are not able to span the information contained in the BS6 and DHS3 models. Adding the BS6 model to the DHS3 model yields an average improvement of 30.5%, whereas vice versa the gain is 22.9%.

The economic gain could also depend on the specific construction choice. Between parentheses, we report the standard deviation of the improvements in Sharpe ratios, across construction methods. For example, on average, the BS6 model improves the FF5 model by 42.4%, but also has a substantial standard deviation of 13.2%. This implies that there is a construction set in the 95% confidence interval for which the improvement is 16.5%, but also a set for which the improvement is 68.3%. The Q-factor model improves the $FF6_c$ by 7.6% on average, with a standard deviation of 6.6%. Hence, in some cases, it may appear that the economic gain is (close to) zero, thereby giving the illusion that the Q4 model is not able to improve the $FF6_c$ model. The main takeaway is that the

Table 4
Economic significance. This table reports the increase in the maximum Sharpe ratio of the augmented model ($M1UM0,t$) relative to the base model ($M0$), to quantify the economic significance: $\Delta\%Sh(M0,M1) = Sh(M0,M1)/Sh(M0) - 1$. The table reports the results using gross returns. The standard deviation of the increase in Sharpe ratio, across construction methods, is reported in parentheses. The data runs from July 1972 until December 2021.

M0	$M1UM0,t$						
	FF5	$FF5_c$	FF6	$FF6_c$	Q4	BS6	DHS3
FF5		18.7 (9.1)	27.3 (10.7)	39.5 (12.8)	31.1 (16.5)	42.4 (13.2)	59.3 (17.6)
$FF5_c$	1.8 (2.7)		19.8 (8.1)	18.7 (7.7)	17.6 (12.8)	27.6 (9.8)	44.8 (16.3)
FF6	0.0 (0.0)	9.6 (4.7)		9.6 (4.7)	12.1 (8.8)	12.1 (8.8)	30.4 (10.3)
$FF6_c$	1.0 (1.4)	0.0 (0.0)	1.0 (1.4)		7.6 (6.6)	7.6 (6.6)	25.7 (10.3)
Q4	3.8 (3.5)	8.6 (7.3)	13.7 (11.3)	18.8 (13.9)		12.1 (10.3)	29.1 (9.9)
BS6	1.3 (1.9)	5.8 (5.0)	1.3 (1.9)	5.8 (5.0)	0.0 (0.0)		22.9 (5.7)
DHS3	20.1 (10.1)	26.9 (10.7)	25.1 (9.6)	30.8 (10.4)	22.6 (12.1)	30.5 (11.7)	

Table 5
In-sample and out-of-sample Sharpe ratios. This table reports the percentage of bootstrap simulations where the maximum Sharpe ratio of the model in the row exceeds that of the model in the column, averaged across construction methodologies. "SR" reports the max. Sharpe ratio of the row model, averaged across construction methodologies. $\sigma(SR)$ reports the standard deviation of the max. Sharpe ratio of the row model. "Best" reports the estimated probability that the row model produces the highest squared Sharpe ratio among all models in the run, averaged over construction methods. $\sigma(Best)$ reports the corresponding standard deviation. Panel A presents the in-sample (IS) estimates and Panel B shows the out-of-sample (OS) estimates, based on 100.000 runs. Each run splits the 594 sample months, running from July 1972 until December 2021, into 297 adjacent pair-months. The run randomly draws a sample of pairs (with replacement). The IS simulation randomly draws one month from each pair within a run. The remaining months form the OS. The IS observations are used to calculate IS Sharpe ratios and portfolio weights. The IS portfolio weights are applied to the OS returns to produce an OS Sharpe ratio estimate.

Panel A: In-sample estimates

	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	5.4	0.0	0.4	14.0	1.2	16.3	0.00	0.00	1.18	0.17
FF6	94.6	0.0	24.3	0.0	38.5	11.1	33.9	0.00	0.00	1.35	0.19
$FF5_c$	100.0	75.7	0.0	10.0	56.7	16.9	49.9	1.26	3.66	1.50	0.25
$FF6_c$	99.6	100.0	90.0	0.0	69.6	36.3	64.8	26.44	30.27	1.61	0.25
Q4	86.0	61.5	43.3	30.4	0.0	0.0	42.1	0.00	0.00	1.44	0.26
BS6	98.8	88.9	83.1	63.7	100.0	0.0	74.6	53.99	39.15	1.76	0.35
DHS	83.7	66.1	50.1	35.2	57.9	25.4	0.0	18.31	19.94	1.50	0.28

Panel B: Out-of-sample estimates

	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	39.7	8.2	9.8	21.6	5.6	22.6	0.43	0.47	0.67	0.11
FF6	60.3	0.0	11.1	10.3	25.9	7.1	24.5	0.86	0.71	0.70	0.12
$FF5_c$	91.8	88.9	0.0	48.8	67.3	25.1	53.5	8.34	7.85	1.00	0.16
$FF6_c$	90.2	89.7	51.2	0.0	67.4	28.0	54.2	10.49	7.75	1.01	0.18
Q4	78.4	74.1	32.7	32.6	0.0	17.6	42.1	8.57	10.61	0.88	0.14
BS6	94.4	92.9	74.9	72.0	82.4	0.0	64.7	41.68	22.32	1.13	0.20
DHS	77.4	75.5	46.5	45.8	57.9	35.3	0.0	29.63	28.80	0.97	0.26

improvement in Sharpe ratio, when adding additional factors, is not only a function of expected returns, variances, and correlations among factors, but also a function of factor construction choices.

6.4. In-sample and out-of-sample estimation

So far, we have used full-sample estimates to calculate maximum Sharpe ratios for our model selection exercise. When factors have high average returns relative to expected returns, these factors obtain too much weight in the ex-post mean–variance tangency portfolio. The optimal mean–variance efficient weights will be overfit, even though they are noisy estimates of the true weights. Consequently, the estimates of the maximum Sharpe ratio can be biased upwards. This bias becomes larger in smaller samples, since the parameter estimates have more sampling error. Also, the bias in the estimates of the maximum Sharpe ratio is especially problematic for comparing non-nested models, such as the Q-factor model versus Fama–French models. To solve this problem, we run bootstrap simulations of in-sample (IS) and out-of-sample (OS) Sharpe ratio estimates, following Fama and French (2018). The

bootstrap approach has the advantage, compared to the full-sample approach, that it is able to yield a distribution of maximum Sharpe ratio estimates and that it allows for testing how often one model outperforms the other.

The bootstrap procedure that we use is to split the 594 months into 297 adjacent pairs of months for a given set of factors constructed from construction rule r. For each simulation run, we draw (with replacement) a random sample of 297 pairs. We randomly assign a month from each pair to the IS sample. Using this IS sample of factor returns, we compute the maximum Sharpe ratio for each model and the corresponding mean–variance optimal portfolio weights. We allocate the remaining unassigned months to the OS sample. Subsequently, we compute the out-of-sample Sharpe ratio estimate using the OS sample of factor returns and the weights estimated from the IS sample. The IS estimates are, like the full-sample estimates, subject to an upward bias. However, this is less of a problem for OS Sharpe ratios, since monthly returns are approximately serially uncorrelated. For each construction rule r we run 100.000 simulation runs. For each run, we compare the maximum Sharpe ratio between models and count how many times a model has a higher maximum Sharpe ratio than an other model. By doing so, we can calculate both the in-sample and out-of-sample probability that a model is winning from other models. In addition, we can calculate this win-probability within simulation r and the total win-probability averaged across all construction rules.

Table 5 shows the win-probability estimates obtained from the bootstrap simulations. Panel A shows the in-sample estimates, which should be interpreted with caution as the in-sample Sharpe ratios are upward biased and based on 297 observation months. We find that the $FF6_c$ model outperforms the Q factor model in 69.6% of the sample. Its Sharpe ratio (1.61) is slightly higher than that of the Q factor model (1.44). The BS6 model seems to outperform the other models, with pairwise win-probabilities over 63.7% and an average Sharpe ratio of 1.76. It is, on average, the model with the highest Sharpe ratio in 54.0% of all simulation runs. The standard deviation is 39.2%, implying large variation across construction methods. The FF6c model in this simulation has an average Sharpe ratio of 1.61, making it the second-best model in this aspect. Still, this model has the largest Sharpe ratio in 26.4% of all simulation runs.

Panel B presents the out-of-sample (OS) results. We find that the BS6 model outperforms all other models in 41.7% of the simulation runs in an out-of-sample setting, averaged across construction methods. The DHS model obtains an overall win-probability of 29.6%, making it the second strongest model from an out-of-sample perspective. Both models have a standard deviation of over 20% in the overall win-probability. This implies that in many construction choices one model may appear superior to the other, and vice versa, while other models, such as the $FF6_c$ model, also have a win-probability exceeding zero. Given the high standard deviations, our conclusion is again that one should be cautious when drawing inferences from one or a few sets of construction choices.

7. Portfolio characteristics across construction choices

We have shown that factor returns vary significantly across different sets of construction choices and that different construction choices can have an influence on model selection exercises. In this section, we study how variation in construction choices affects portfolio characteristics that, in turn, have an impact on portfolio performance. We consider the factor exposure, illiquidity and transaction costs of a portfolio.

Regarding factor exposure, the expected return of a well-diversified factor portfolio is directly related to the sorting characteristic (Cochrane, 2011). We define factor exposure by creating a normalized factor score. For every variable v, we first compute the cross-sectional average, maximum and minimum at time t. Next, for every stock i, we compute the normalized factor score for all variables v at time t by subtracting the cross-sectional average from the variable score of the stock $variable_{i,v}$ and subsequently dividing by the spread between the maximum variable score in that month and the minimum variable score in that month[13]:

$$Normalized\ factor\ score_{i,v,t} = \frac{Variable_{i,v,t} - Mean_{v,t}}{Max_{v,t} - Min_{v,t}} \tag{4}$$

In both the long and the short side of the long–short portfolio, we aggregate the normalized factor scores to the portfolio level by using the respective construction choices. Subsequently, we compute the spread between the long and short leg of the factor portfolio to arrive at the factor exposure per factor per construction choice.

In addition, construction choices may impact the liquidity of a portfolio. Stocks with low liquidity, such as microcaps, may have high transaction costs and other frictions (such as relatively high bid–ask spreads), which could directly impact the returns of factor portfolios. We measure the liquidity of the portfolios by aggregating stock-level illiquidity to portfolio-level illiquidity following Amihud (2002). More specifically, we measure stock-level illiquidity as the average ratio of the daily absolute return to the dollar trading volume on month t:

$$ILLIQ_{i,t} = \frac{1}{D_{i,t}} \sum_{t=1}^{D_{i,t}} \frac{|R_{i,t,d}|}{VOLD_{i,t,d}} \tag{5}$$

The daily return of a stock is denoted by $R_{i,t,d}$. $VOL_{i,t,d} * P$ equals the dollar trading volume for stock i on day d of month t. $D_{i,t}$ equals the amount of trading days for stock i on month t. A lower value of $ILLIQ_{i,t}$ implies a higher level of liquidity.

We further consider whether construction choices affect transaction costs. We estimate transaction costs at the individual stock-level using the procedure of Hasbrouck (2009). This procedure allows us to estimate effective spreads for individual stocks using

[13] We calculate the normalized factor score before using exclusion criteria.

Table 6
Portfolio characteristics. This table shows the estimated coefficients obtained from fixed effect regressions about the relation between eleven construction choices and ex-ante long–short normalized factor exposure, portfolio illiquidity, and transaction costs. The construction choice definitions are the same as in Fig. 3. The normalized factor exposures are calculated for each firm on a monthly basis and aggregated to a portfolio level. The normalized firm factor exposure is calculated as: $(Variable - Mean)/(Max - -Min)$. Illiquidity is calculated following Amihud (2002) and transaction costs following Hasbrouck (2009). Monthly characteristics are constructed using data from July 1972 to December 2021. Factor fixed effects and time fixed effects are included. Observations are weighted by factor. Double-clustered (by factor and date) adjusted t-statistics are reported between parentheses (Thompson, 2011).

	30–70	NYSE	BE	Micro	PRC	Utilities	Financials	Industry	Independent	Rebal	23
Factor-	−1.17***	−0.14	−1.18***	−0.59***	−1.45***	−0.05	−0.15**	−0.23***	0.01	0.03	−0.05
Exposure	(−4.43)	(−1.05)	(−4.73)	(−4.24)	(−4.65)	(−1.63)	(−2.65)	(−3.73)	(0.98)	(1.25)	(−1.55)
Illiquidity	−0.12**	−0.84***	−0.05**	2.04***	−1.87***	−0.10***	0.03*	−0.01	−0.03	−0.01	−0.08*
	(−2.35)	(−8.53)	(−2.53)	(5.75)	(−4.89)	(−8.12)	(1.95)	(−0.49)	(−0.90)	(−0.12)	(−1.75)
Transaction-	−0.05***	−0.14***	−0.00	0.09***	−0.08***	−0.01***	−0.01***	0.04***	−0.00***	−0.01***	−0.02***
Cost	(−5.25)	(−6.75)	(−0.93)	(6.73)	(−4.48)	(−8.11)	(−9.59)	(5.83)	(−3.72)	(−4.33)	(−5.38)

* indicate significance level at 10%.
** indicate significance level at 5%.
*** indicate significance level at 1%.

their daily price series. To examine the impact of construction choices on factor exposure, illiquidity and transaction costs, we run fixed-effect panel regressions where we regress the constructed variables on dummy variables of each construction choice. We include factor and time fixed effects in the estimation. Table 6 shows the estimated coefficients.

Overall, most construction choices significantly impact portfolio characteristics. We find that 6, 6 and 10 out of the 11 construction choices show significant coefficients (at the 5% level) on factor exposure, portfolio illiquidity, and transaction costs, respectively. Portfolios based on 30–70 breakpoints have significantly lower factor exposures than those with 20–80 breakpoints, while they are more liquid. Furthermore, 30–70 portfolios have, on average, 5 basis points lower transaction costs than 20–80 breakpoints portfolios. Using NYSE instead of NAN breakpoints significantly lowers transaction costs by an average of 14 basis points and improves portfolio liquidity. This is intuitive as NAN breakpoints allow more small firms to enter the portfolio, hence increasing transaction costs and illiquidity. Excluding stocks with a price below 5 dollars has a significant negative impact on factor exposures, while at the same time improving liquidity and reducing transaction costs. Including financial firms and utility firms also reduces transaction costs, albeit only with 1 basis point. Furthermore, we find that the monthly rebalancing frequency and using $2 \times 3 \times 3$ sorts raises portfolio-level transaction costs compared to annual rebalancing and 2×3 sorts, respectively.

8. Conclusion

Within empirical asset pricing, character-based sorting is a popular way to construct asset pricing factors. This paper stresses that constructing factors involves a large number of choices, leading to "degrees of freedom" for researchers. Especially since there is no consensus on construction methods, the degrees of freedom involved allows for p-hacking if the choices affect outcomes: researchers could then pick construction choices in such a way that the resulting factor meet certain statistical and performance-related hurdles, such as high Sharpe ratios.

We describe the different choices made in factor models, and in the wider literature. We find that construction choices impact factor returns. Using 2048 different combinations of choices, we show large and significant variation in Sharpe ratios based on factor returns. As such, the variation in choices for factor construction by researchers leads to substantial variation in outcomes. We calculate non-standard errors as the standard deviation of the generated Sharpe ratios and show that the non-standard errors in our setting are sizable, also in comparison with standard errors. The variation that we document materially impacts model selection exercises when comparing factor models. Maximum Sharpe ratios of factor models show wide variation across construction methods and also mean–variance weights vary substantially across construction methods. By following a bootstrapping approach, we show that design choices substantially affect a model's probability of producing the highest Sharpe ratio. Our analysis indicates that factor models should not be compared against each other when their construction method differ and that it is important to check how the best-performing model depends on the construction choices being made.

Our results facilitate the selection of the most informative robustness tests. Currently, a typical robustness check examines robustness to the weighting choice. Our results suggest that the other most important design choices around factor construction are those concerning NYSE or NAN breakpoints, micro stocks, industry-adjusted characteristics, and rebalancing frequency. In a specification check (Brodeur et al., 2020, Mitton, 2022), researchers could graphically show the distribution of their Sharpe ratios (or other results) if their design choices are varied among these dimensions.

It is important to again stress the tradeoff involved. Variation in design choices allows researchers to customize samples, or to include the whole market, to test specific research questions. In addition, variation reduces the chance of missed discoveries, giving researchers a better chance to push the frontier. However, too much variation severely complicates the comparison of results across papers. Our message is that authors should clearly describe their design choices, and should explain deviations from the norm. By keeping a very limited set of construction choices fixed, empirical asset pricing researchers can greatly facilitate the interpretability of their presented results, as many of the design choices that we study have more moderate effects. Stressing this more comforting message to the empirical asset pricing field is also important.

CRediT authorship contribution statement

Amar Soebhag: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bart Van Vliet:** Writing – review & editing, Writing – original draft, Formal analysis. **Patrick Verwijmeren:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Conceptualization.

Appendix. Sorting variables

This Appendix explains the sorting variables in more detail.

Market: Value-weighted return on the CRSP stock market index minus the risk-free rate.

Market capitalization: The price (CRSP item PRC) times shares outstanding (CRSP item SHROUT). Market capitalization is used to construct the size factor (SMB).

Book-to-market ratio: Book equity in the sort for June of year t is defined as the total assets for the previous fiscal year-end in calendar year $t-1$, minus liabilities, plus deferred taxes and investment tax credit, minus preferred stock liquidating value if available or redemption value if available, or carrying value. The carrying value is adjusted for net share issuance from the fiscal year-end to the end of December of $t-1$. Market capitalization is price times shares outstanding at the end of December of $t-1$, from CRSP. The book-to-market ratio is used to construct the value factor (HML).

Growth in book assets: Growth in book assets, in year t, is defined as the change in total assets from the fiscal year ending in $t-2$ to the fiscal year ending in $t-1$ divided by total assets at $t-2$. This signal is used to construct the CMA and IA factor.

Operating Profitability: Operating Profitability in the sort for June of year t is measured with accounting data for the fiscal year ending in year $t-1$ and is revenues minus cost of goods sold, minus selling, general, and administrative expenses, minus interest expense, minus research and development expenses, all divided by book equity. This signal is used to construct the RMW factor.

Cash Profitability: The operating profitability minus accruals for the fiscal year ending in $t-1$. Accruals are the change in accounts receivable from $t-2$ to $t-1$, plus the change in prepaid expenses, minus the change in accounts payable, inventory, deferred revenue, and accrued expenses (Ball et al., 2016). This signal is used to construct the cash-based RMW factor.

Momentum: The cumulative return between month $t-12$ and $t-2$, which is used to construct the UMD factor.

Return-on-equity: Income before extraordinary items (Compustat quarterly item IBQ) divided by 1-quarter-lagged book equity. Earnings data are used in the months immediately after the most recent public quarterly earnings announcement dates (Compustat item RDQ). In addition, we require the end of the fiscal quarter that corresponds to its most recently announced quarterly earnings to be within 6 months prior to the portfolio formation, to exclude stale earnings. We use this signal to construct the ROE factor.

Composite share issuance: The composite share issuance is the firm's 5-year growth in market equity, minus the 5-year equity return, in logs. We use this signal, together with net share issuance, to construct the financing (FIN) factor.

Net share issuance: This signal is similar to the composite share issuance, except that we use a 1-year horizon and exclude cash dividends.

Cumulative abnormal returns earnings announcement: We compute the cumulative abnormal returns around earnings announcements as the 4-day cumulative abnormal return from day $t-2$ to $t+1$ around the latest quarterly earnings announcement date (Compustat item RDQ):

$$CAR_i = \sum_{d=-2}^{d+1} (R_{i,d} - Rm, d) \tag{6}$$

$R_{i,d}$ denotes the stock return on day d and $R_{m,d}$ denotes the market return. We use the cumulative abnormal return in the months immediately following the quarterly earnings announcement date, but within 6 months from the fiscal quarter end (to exclude stale earnings). We require earnings announcement dates to be after the corresponding fiscal quarter end. In addition, we require valid daily returns on at least two of the trading days in the CAR window. We require the Compustat earnings date (RDQ) to be at least two trading days prior to the month end. We use the most recent CAR to construct the PEAD factor.

References

Ahmed, S., Bu, Z., Tsvetanov, D., 2019. Best of the best: A comparison of factor models. J. Financ. Quant. Anal. 54 (4), 1713–1758.

Akey, P., Robertson, A.Z., Simutin, M., 2022. Noisy factors. Available at SSRN 3959116.

Amihud, Y., 2002. Illiquidity and stock returns: Cross-section and time-series effects. J. Financial Mark. 5 (1), 31–56.

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. J. Finance 61 (1), 259–299.

Bali, T.G., Brown, S.J., Tang, Y., 2017. Is economic uncertainty priced in the cross-section of stock returns? J. Financ. Econ. 126 (3), 471–489.

Ball, R., Gerakos, J., Linnainmaa, J.T., Nikolaev, V., 2016. Accruals, cash flows, and operating profitability in the cross section of stock returns. J. Financ. Econ. 121 (1), 28–45.

Barillas, F., Kan, R., Robotti, C., Shanken, J., 2020. Model comparison with sharpe ratios. J. Financ. Quant. Anal. 55 (6), 1840–1874.

Barillas, F., Shanken, J., 2017. Which alpha? Rev. Financ. Stud. 30 (4), 1316–1338.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. J. Finance 73 (2), 715–754.

Brodeur, A., Cook, N., Heyes, A., 2020. Methods matter: p-hacking and publication bias in causal analysis in economics. Amer. Econ. Rev. 110 (11), 3634–3660.

Brown, S.J., Lajbcygier, P., Li, B., 2008. Going negative: What to do with negative book equity stocks. J. Portf. Manag. 35 (1), 95–102.

Chen, A.Y., Zimmermann, T., 2022. Open source cross-sectional asset pricing. Crit. Finance Rev. 11 (2), 207–264.

Cochrane, J.H., 2011. Presidential address: Discount rates. J. Finance 66 (4), 1047–1108.

Daniel, K., Hirshleifer, D., Sun, L., 2020a. Short-and long-horizon behavioral factors. Rev. Financ. Stud. 33 (4), 1673–1736.

Daniel, K., Mota, L., Rottke, S., Santos, T., 2020b. The cross-section of risk and returns. Rev. Financ. Stud. 33 (5), 1927–1979.

Detzel, A., Novy-Marx, R., Velikov, M., 2023. Model comparison with transaction costs. J. Finance.
Drechsler, I., Drechsler, Q.F., 2014. The Shorting Premium and Asset Pricing Anomalies. Tech. Rep., National Bureau of Economic Research.
Ehsani, S., Harvey, C.R., Li, F., 2021. Is sector-neutrality in factor investing a mistake? Available at SSRN 3930228.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. J. Finance 47 (2), 427–465.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.
Fama, E.F., French, K.R., 2008. Dissecting anomalies. J. Finance 63 (4), 1653–1678.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Fama, E.F., French, K.R., 2018. Choosing factors. J. Financ. Econ. 128 (2), 234–252.
Foerster, S.R., Sapp, S.G., 2005. Valuation of financial versus non-financial firms: A global perspective. J. Int. Financ. Mark., Inst. Money 15 (1), 1–20.
Gibbons, M.R., Ross, S.A., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 1121–1152.
Harvey, C.R., Liu, Y., Zhu, H., 2016. . . . and the cross-section of expected returns. Rev. Financ. Stud. 29 (1), 5–68.
Hasbrouck, J., 2009. Trading costs and returns for US equities: Estimating effective costs from daily data. J. Finance 64 (3), 1445–1477.
Hasler, M., 2023. Looking under the hood of data-mining. Available at SSRN.
Hasler, M., 2024. Is the value premium smaller than we thought? Crit. Finance Rev. (forthcoming).
Hirshleifer, D., Hsu, P.-H., Li, D., 2018. Innovative originality, profitability, and stock returns. Rev. Financ. Stud. 31 (7), 2553–2605.
Hou, K., Mo, H., Xue, C., Zhang, L., 2019. Which factors? Rev. Finance 23 (1), 1–35.
Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Rev. Financ. Stud. 28 (3), 650–705.
Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Rev. Financ. Stud. 33 (5), 2019–2133.
Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2023. Is there a replication crisis in finance? J. Finance 78 (5), 2465–2518.
Kessler, S., Scherer, B., Harries, J.P., 2020. Value by design? J. Portf. Manag. 46 (2), 1–19.
Linnainmaa, J.T., Roberts, M.R., 2018. The history of the cross-section of stock returns. Rev. Financ. Stud. 31 (7), 2606–2649.
McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Finance 71 (1), 5–32.
Menkveld, A.J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüss, S., Razen, M., et al., 2023. Non-standard errors. J. Finance 79 (3), 2339–2390.
Mitton, T., 2022. Methodological variation in empirical corporate finance. Rev. Financ. Stud. 35, 527–575.
Novy-Marx, R., 2013. The other side of value: The gross profitability premium. J. Financ. Econ. 108 (1), 1–28.
Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. Rev. Financ. Stud. 29 (1), 104–147.
Stambaugh, R.F., Yuan, Y., 2017. Mispricing factors. Rev. Financ. Stud. 30 (4), 1270–1315.
Thompson, S.B., 2011. Simple formulas for standard errors that cluster by both firm and time. J. Financ. Econ. 99 (1), 1–10.
Wahal, S., Yavuz, M.D., 2013. Style investing, comovement and return predictability. J. Financ. Econ. 107 (1), 136–154.
Walter, D., Weber, R., Weiss, P., 2023. Non-standard errors in portfolio sorts. Available at SSRN.
Yan, X.S., Zheng, L., 2017. Fundamental analysis and the cross-section of stock returns: A data-mining approach. Rev. Financ. Stud. 30 (4), 1382–1423.