Non-Standard Errors in Asset Pricing: Mind Your Sorts

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

[a]Erasmus School of Economics, the Netherlands
[b]Robeco Quantitative Investing, the Netherlands
[c]University of Melbourne, Australia

August, 2022

Abstract

Non-standard errors capture uncertainty due to differences in research design choices. We establish substantial variation in the design choices made by researchers when constructing asset pricing factors. By purposely data mining over two thousand different versions of each factor, we find that Sharpe ratios exhibit substantial variation within a factor due to different construction choices, which results in sizable non-standard errors and allows for p-hacking. We provide simple suggestions that reduce the average non-standard error by 70%. Our study has important implications for model selection exercises.

JEL Classification: G11, G12, G15
Keywords: non-standard errors, portfolio construction, factor investing, equity factors, asset pricing models, p-hacking, data-mining.

A. Soebhag: soebhag@ese.eur.nl (corresponding author: Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3000 DR, Netherlands), B.P. van Vliet: b.p.vanvliet@ese.eur.nl, P. Verwijmeren: verwijmeren@ese.eur.nl. We thank Guido Baltussen, Paul Fontanier, Peter Koudijs, Frederik Muskens, Dimitris Papadimitriou, Esad Smaljbegovic, Sjoerd van Bekkum, seminar participants at Robeco Institutional Asset Management, and conference participants at the 2022 Finance Symposium for fruitful discussions and feedback. The views expressed in this paper are not necessarily shared by Robeco Institutional Asset Management. Declarations of interest: none.

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). Interestingly, the construction procedure is not uniform across studies. In fact, researchers face a number of design choices when engaging in portfolio sorting and the choices being made vary substantially across studies. These differential choices lead to non-standard errors (Menkveld et al., 2021): in addition to a data-generating process, there is an evidence-generating process that leads to uncertainty due to variation in research design choices.

In this paper, we study the extent to which the differential design choices in portfolio sorting matter for factors and factor models. A better understanding of the design choices that matter allows researchers to more effectively show the robustness of their findings in future work, while also facilitating model selection exercises and helping interested readers in interpreting presented results. Understanding the non-standard errors in asset pricing is also important because researchers may have incentives to engage in p-hacking (Harvey, 2017): ambiguity regarding construction choices creates room for researchers to construct factors in a way that maximizes some statistical criteria, such as maximizing Sharpe ratios and t-statistics.

Based on a review of a representative set of 323 empirical asset pricing studies originating from Harvey and Liu (2019), we consider eleven construction choices that researchers face in their research design. These choices are: (1) 70/30 or 80/20 breakpoints, (2) NYSE or NYSE-AMEX-Nasdaq (NAN) breakpoints, (3) including or excluding firms with a negative book equity value, (4) including or excluding microcaps, (5) imposing a price filter or not, (6) including or excluding utility firms, (7) including or excluding financial firms, (8) industry neutralization or not, (9) value-weighting or equal-weighting, (10) independent or dependent sorts, and (11) sorting on the most recent market capitalization or from June. We construct factors using each possible combination of choices, which leads to 2048 (2^{11}) construction combinations.

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 data on U.S. stock returns from January 1972 to December 2021, we find that factors exhibit large variation in Sharpe ratios within our set of possible construction methods. As an illustration, Figure 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 0.49. 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 0.15 and 1.24. Our paper shows that the same design choices can also strongly affect the Sharpe ratios of other factors.

The non-standard errors in our setting can be defined as the standard deviation of the generated Sharpe ratios across the possible construction methods. We find that these non-standard errors are sizable relative to standard errors, across all factors. In multiple cases, the non-standard errors exceed the standard errors. For example, the non-standard error for the post-earnings announcement drift factor is 0.10, whereas the standard error ranges between 0.04 and 0.09. The average ratio of the non-standard error to the standard error across factors is 1.18. As such, factor returns are not only a function of their sorting characteristic, but also a function of their construction choices.

The above non-standard errors are based on researchers randomly choosing construction methods, which helps in assessing the room provided by these choices to optimize along a given criteria. An alternative calculation of non-standard errors takes into account that not all choices are equally likely, as researchers could have good reasons to select a particular choice, or are simply more likely to select a choice that they have seen more often in earlier work. For our alternative non-standard error measure, we classify the choices made in the set of 323 empirical asset pricing papers originating from Harvey and Liu (2019). We exploit the popularity of each binary choice in earlier work to construct non-standard errors that take these probabilities into account. Interestingly, no binary option is so dominant that it represents close to 100% of the observed choices, and some options are even selected with a roughly 50% probability. Using these probabilities, we find that non-standard errors remain sizable, with an average ratio of non-standard errors to standard errors of 1.08.

Factor models have been compared against each other based on factor construction methodology choices of a single paper. Our paper aims to perform a model comparison without relying on a single set of construction choices but by considering a wide set of potential choices. 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] We find that the factor models of Barillas and Shanken (2018) and Daniel, Hirshleifer, and Sun (2020) have the largest maximum Sharpe ratio. Importantly, this maximum Sharpe ratio, and with it the dominant factor model, also varies across construction methods.

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

[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.

transaction costs of a portfolio. We again find that portfolio construction methods matter. For example, equal-weighting leads to portfolios with higher illiquidity than value-weighting, which consequently results in higher risk-adjusted gross returns.

Overall, we conclude that factor design choices matter. Our results imply that multiple construction methods should be considered to reduce the potential for data mining. We find that particularly important choices are those concerning the use of NYSE or NAN breakpoints, including or excluding micro stocks, industry-adjusted characteristics, and equal versus value-weighting. For future studies, one 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 these methodological possibilities are reported. Another way forward is for studies to be more uniform in their choices. Based on our analysis, the conservative choices of using NYSE breakpoints, excluding microcaps, and using value-weighting reduces the average non-standard error by 70%.

Our paper contributes to empirical studies on the replicability of market anomalies. 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.[2] McLean and Pontiff (2016) test anomalies out-of-sample and find that the performance of identified anomalies diminishes after publication. Linnainmaa and Roberts (2018) find that a similar conclusion can be drawn when examining pre-sample periods. 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. Hou et al. (2020) test 452 anomalies by using a single

[2] Methods to control for the multiple testing problem are further developed in Harvey and Liu (2020) and Giglio et al. (2021). Kozak et al. (2020) focus on assessing the joint pricing role of a large number of factors and the potential redundancy among the candidate factors. Feng et al. (2020) provide a framework for systematically evaluating the contribution of individual factors relative to existing factors and for conducting statistical inference in such a high-dimensional setting. Li and Robotti (2022) propose a simulated-based approach to benchmark the evidence in asset pricing tests.

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. In our study, we also compare factor models on an "apples-to-apples" basis, but we start with an agnostic view about construction choices and consider a wide range of construction sets, as our goal is to quantify the importance of a range of construction choices.

Our set of construction choices includes those questioned by earlier studies. In studies where data are scarce or in an international empirical asset pricing context (e.g., Ang et al. (2006), Novy-Marx (2013), Wahal and Yavuz (2013)), the use of dependent sorts has been advocated instead of independent sorts, as the latter could produce empty portfolios. Kessler et al. (2020) consider choices such as how to determine investment weights when targeting an audience interested in the practical aspects of implementing a value strategy. Hou et al. (2019) show that the performance of factors is sensitive to the breakpoints being used, whereas Hou et al. (2020) show that many anomalies disappear when microcaps are excluded. We look at the combined set of construction choices to get a better idea of the overall importance of these choices and the magnitude of non-standard errors. Non-standard errors are introduced by Menkveld et al. (2021), who argue that a layer of uncertainty in academic work is due to the evidence-generating process that exposes variation across choices by researchers, in addition to the traditional standard errors following from the uncertainty in sample estimates of population parameters. By letting 164 research teams independently test the same market microstructure hypotheses on the same sample of trade records, they find that non-standard errors are sizable. Our paper can be thought of as modelling N hypothetical researchers who construct factor returns. Our approach allows us to examine non-standard errors both when researchers freely exploit the variation of choices available to them, and when researchers base their choices on conventions in earlier work. We find relatively high non-standard errors in both scenarios.

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. In Section 4 we examine the importance of factor construction choices for Sharpe ratios and calculate non-standard errors. Section 5 examines whether factor construction choices impact model selection exercises. Section 6 shows how different construction methods affects several key portfolio characteristics. Section 7 provides recommendations to reduce non-standard errors. 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 and with share codes 10 or 11, which limits our sample to common stocks. The sample period spans January 1972 to December 2021, thereby covering 600 months of factor returns.[3]

We use multiple factor models, originating from Fama and French (2015), Hou et al. (2015), Fama and French (2018), Barillas and Shanken (2018), and Daniel, Hirshleifer, and Sun (2020). Table 1 summarizes the factors underlying the factor models and their key construction choices as used in their original studies.[4] The market factor is a part of all models. The Fama-French 5-factor model of Fama and French (2015) (FF5) consists of the market, size (SMB), value (HML), profitability (RMW) and investment (CMA) factors. The factors are constructed, originally, by using a 2 by 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, both on the NYSE universe. All factors of the Fama-French

[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 Appendix A.

5 factor model are rebalanced yearly. The 6-factor model of Fama and French augments the 5-factor model by adding the momentum (UMD) factor, with the resulting model being abbreviated as FF6. 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 as suggested by 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 factor, size factor, investment (IA) factor, and return on equity (ROE) factor. In the original set-up, these factors are derived from a 2x3x3 independent sort.[5] 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 factor, and the return on equity factor. Daniel, Hirshleifer, and Sun (2020) (DHS) construct a three-factor model consisting of the market factor, financing (FIN) factor, and the post-earnings announcement drift (PEAD) factor. Both the FIN and PEAD factor use 20-80 breakpoints in the characteristic dimension. The PEAD factor is rebalanced monthly.

We construct factor portfolios by sorting on both market capitalization and a factor characteristic. The size dimension is split into a "Small" and a "Big" segment based on the median. The characteristic dimension is split into a "Low", "Neutral", and "High" portfolio based on two breakpoints.[6] This procedure, the 2x3 sorting, results into six portfolios: Small.Low, Small.Neutral, Small.High, Big.Low, Big.Neutral and Big.High. We create the factor portfolio by taking a long position in the Small.High and the Big.High portfolio and a short position in the Small.Low and Big.Low portfolio:

$$Factor = (Small.High + Big.High)/2 - (Small.Low + Big.Low)/2 \tag{1}$$

[5] We examine the 2x3x3 versus the 2x3 sorting procedure in Appendix B.

[6] For the financing factor we follow the approach in Daniel, Hirshleifer, and Sun (2020) by separately sorting all repurchasing firms into two groups using a median breakpoint, and sorting all issuing firms into three groups using two breakpoints.

3 Variation in Sorting Methods

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 survey 323 empirical articles in the top finance journals between 1965 and 2018, based on the list of papers that Harvey and Liu (2019) constructed for their census of the "factor zoo". In this section, we start by describing the choices that these studies face. We then document how often certain choices are being made.

3.1 Construction choices

Based on the data and methodology sections of the 323 empirical asset pricing studies, we find that there are eleven construction choices that are commonly being mentioned.[7] These eleven choices result in a set of 2048 (2^{11}) construction choices, which translates to 2048 different versions for each factor and factor model. In this subsection, we explain the eleven choices one by one.

3.1.1 Characteristic breakpoints

Common practice in the academic finance literature has been to create portfolios by sorting on characteristics positively 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

[7] 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).

70th percentile (calculated over the NYSE universe) of the characteristic. The intersection of these classifications results into six portfolios, from which the high-minus-low portfolio is derived.

The 30th and 70th percentile breakpoints are thus one popular choice, used in, for example, 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. Examples of studies using this method are McLean and Pontiff (2016), Stambaugh and Yuan (2017) and Daniel, Hirshleifer, and Sun (2020). The consequence of using the latter choice is that stocks with more extreme characteristics are selected into portfolios. We construct different versions of factors where we either use the 30th-70th breakpoint or the 20th-80th breakpoint in the characteristic dimension.

3.1.2 Breakpoints universe

A common choice is to calculate breakpoints over the NYSE universe. However, a popular alternative is to calculate breakpoints over the NYSE-AMEX-Nasdaq (NAN) universe, such as done by McLean and Pontiff (2016), Stambaugh and Yuan (2017) and Yan and Zheng (2017). Since Nasdaq and AMEX stocks have a tilt towards smaller stocks, the median market capitalization is always higher under the NYSE criteria relative to the NAN criteria. As such, using NAN breakpoints is likely to provide an overweight towards micro- and small-cap stocks relative to using NYSE breakpoints.

3.1.3 Negative book equity value

Firms with a negative book equity value were rare before 1980 (Fama and French (1993)). However, they represent a larger proportion of firms over time, even though negative book equity has no obvious interpretation due to a firm's limited liability structure (Brown et al. (2008)). Many practitioners and academics omit negative book equity firms from their analysis, but an even larger set of papers still contains analyses with negative book equity value firms included. Brown et al. (2008) show that negative book equity firms are disproportionately represented in extreme growth and value sectors.

3.1.4 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 high transaction costs (Novy-Marx and Velikov (2016)). Nevertheless, microcaps are typically included in many studies. Hou et al. (2020) find that many anomalies documented in the literature do not survive after excluding microcaps. Excluding microcaps increases the median market capitalization, reduces typical return volatility, and increases the market share of stocks below the median.

3.1.5 Filtering on price

A price filter leads researchers to exclude firms solely based on absolute share prices. More specifically, stocks are dropped for having share prices below a minimum, which typically varies between $1 and $5. In fact, the most common price filters use a minimum of exactly $1 (e.g, Lee and Swaminathan (2000)) or exactly $5 (e.g, Amihud (2002)). Applying a price filter removes potentially highly illiquid and often highly volatile stocks.

3.1.6 Utility firms

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. In empirical corporate finance studies, it is standard to exclude utility firms from the analysis, as they are seen as different due to the regulations utility firms have to comply with. These regulations could also explain the exclusion of utility firms in asset pricing studies, such as in Hirshleifer and Jiang (2010), who argue that mispricing is more constrained among regulated industries. Still, most empirical asset pricing studies incorporate utility firms in their analysis.

3.1.7 Financial firms

Excluding financial firms from the sample is not unusual in empirical studies. The argument for this exclusion criteria is that financial services are fundamentally different, resembling the potential argument for utility firms. Fama and French (1992) explicitly mention that financial firms have high leverage, which is normal for such firms, and that it probably does not have the same meaning as for non-financial firms, where high leverage is more likely to indicate distress. Still, many other papers include financial firms, such as Stambaugh and Yuan (2017). Including financial firms may especially impact factor returns when factors are not hedged against industry exposure.

3.1.8 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 over-weights 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, Mota, Rottke, and Santos (2020) 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.[8] A comparison of the standard and industry-hedged factors shows that industry adjustment often improves factor performance (Asness et al. (2000), Novy-Marx (2013)).

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

$$S^*_{i,t} = (S_{i,t} - \bar{S}_{i,j,t})/(S_{max,j,t} - S_{min,j,t}) \tag{2}$$

$S_{i,t}$ ($S^*_{i,t}$) denotes the (industry-adjusted) sorting characteristic. $\bar{S}_{i,j,t}$, $S_{max,j,t}$ and $\mathrm{S}_{min,j,t}$ 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.

3.1.9 Value-weighting vs. equal-weighting

There are several weighting schemes that a researcher can select when constructing a portfolio. The literature focuses predominantly on value- or equal-weighting portfolios. Different choices regarding weights results in different portfolio compositions and consequently in differential portfolio characteristics and performance. When using the value-weighting approach, these exposures depend on the size of the specific companies. The risk and return

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

will be driven predominantly by the largest companies in the investment universe. Value-weighted portfolios typically serve as a benchmark against which portfolio managers are evaluated, highlighting the relevance of value-weighting in practice. Nevertheless, many studies use equal-weighting when constructing factor portfolios (Hou et al., 2020). In robustness tests, it is common for papers to show the results when the alternative weighting choice would have been selected.

3.1.10 Independent versus dependent

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 2x3 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 2x3x3 dependent sort, it is not clear what the ordering should be, allowing for a wider playing field.

3.1.11 When to observe market capitalization

Common practice is to construct size-breakpoints based on the market capitalization of firms at the end of June of the current year t, and update this yearly, following Fama and French (1992). Some studies have chosen to use the market capitalization in the previous month in their size sort. For example, Daniel, Hirshleifer, and Sun (2020) do so when constructing the PEAD factor, and Ang et al. (2006) in their analysis of the idiosyncratic volatility anomaly.

One argument in favor of using the most recent market capitalization might be to use timely information to construct the size sort. On the other hand, this may result into more turnover, since one rebalances the size sorts each month instead of each year.

3.2 Distribution and correlation of choices

We continue our meta-analysis in this section by documenting how often certain choices are being made, again based on the list of 323 empirical articles constructed for Harvey and Liu (2019). If choices are being made randomly, the proportion of studies in which a particular option is selected will be close to 50% for each of our binary choices. However, if there are good reasons for particular design choices, or if authors build on the choices being made in earlier work, then we might expect some choice options to be selected (close to) 100% of the time. We show the percentage of studies in which a particular design option is selected in Table 2.

For some design choices the distribution is relatively equal. The number of studies reporting to use 30-70 breakpoints roughly equals the number of studies reporting to use 20-80 breakpoints. NYSE breakpoints are used by 41.5% of the studies reporting this information. Value-weighting returns (58.5%) is slightly more popular than equal-weighting returns (41.5%).

No design option is selected in 100% of the cases. The most popular design option is to include utilities, as utility firms are excluded from the sample in only 9.9% of the relevant studies. Financial firms are excluded in 28.8% of the cases. Other popular options are to not impose industry neutrality (88.5%), to include microcaps (88.2%), to not impose a price filter (81.7%), and to include firms with negative book equity values (78.0%). We further find that the proportion of studies using independent sorts is 71.8% and that in 67.4% of the studies the market capitalization from last June is used.

The percentages reported in Table 2 indicate that no design option in our set is extremely rare. Regardless of the choice being made, a researcher can always cite at least ten other papers making the same choice. To examine whether there are combinations of choice options that are particularly rare, we report the correlation matrix of the eleven choices in Table 3.

The typical correlation coefficient is not particularly high. If one, for example, includes microcaps, there is an increased probability of using value-weighting rather than equal-weighting (correlation coefficient of 0.22), but the correlation is not so strong that a choice for equal-weighting would be considered as exceptional. The highest correlations are observed between using independent sorting and using the size from June (0.54) and between excluding firms with negative book equity values and including financial firms (-0.53).

4 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. We focus on Sharpe ratios as these allow us to assess both individual factors in this section, and factor models in the next section (Barillas and Shanken (2017), Fama and French (2018)). In addition, in this section we compute non-standard errors and compare these with estimated standard errors. This section concludes with a subsection on multiple hypothesis testing.

4.1 Construction choices and Sharpe ratios

Table 4 reports summary statistics of the factors that we include in our sample, based on the factor models from Table 1. These factors are the size (SMB), value (HML and the monthly version, HML(m)), operating-based profitability (RMW), cash-based profitability (RMW(cp)), investment (IA and CMA), momentum (UMD), return on equity (ROE), financing (FIN), and post-earnings announcement drift (PEAD) factor. The table shows the annualized average return and Sharpe ratio per factor, both value-weighted and equally-weighted, when averaged over the set of construction methods. Value-weighted factor returns range between 1.91% (SMB) and 8.00% (UMD) per year, with Sharpe ratios ranging between 0.18 (SMB) and 1.10 (PEAD). Returns and Sharpe ratios for equal-weighted factors tend to be higher, except for the size factor.

Figure 2 shows the Sharpe ratio distribution across sets of construction choices for each factor, based on long-short factor returns. We construct a factor 2048 times by using the 2048 different factor construction methods. Figure 2A shows the distribution of value-weighted Sharpe ratios and Figure 2B shows the distribution of equal-weighted Sharpe ratios. Both figures show differences in median Sharpe ratios across factors, but also substantial variation in Sharpe ratios within a factor. For example, based on value-weighting, the Sharpe ratio of the CMA factor ranges between 0.18 and 0.90, the Sharpe ratio of the UMD factor ranges between 0.37 and 0.78, and the Sharpe ratio of the ROE factor ranges between 0.46 and 1.06. Based on equal-weighting, the ranges are between 0.34 and 1.44 for the CMA factor, between 0.28 and 0.92 for the UMD factor, and between 0.42 and 1.40 for the ROE factor. Hence, in relative terms, the Sharpe ratio can more than double 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 0.73 to 1.76 for value-weighted Sharpe ratios and from 0.67 to 2.18 for equal-weighted Sharpe ratios. Overall, these results imply that construction choices matter.

We next examine how specific construction choices, in isolation, affect maximum Sharpe ratio estimates. Figure 3 shows annualized maximum Sharpe ratios by construction choice, averaged over factor models. We first vary the breakpoints that are used to classify high and low characteristics. The first two bars on the left-hand side use the 20th-80th percentile (white bar) or the 30th-70th percentile (dashed bar). The latter case yields an average annualized Sharpe ratio of 0.63, whereas 20-80 breakpoints yield a Sharpe ratio of 0.65.

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 0.55 to 0.73, which is the largest increase within our set of choices. This choice thus comes out as important, where NYSE breakpoints represent the conservative choice. Anoter important choice is the choice whether to include microcaps or not. Including microcaps improves the average Sharpe ratio from 0.59 to 0.68, which makes excluding these firms the conservative choice.

Choices that do not lead to substantially different average Sharpe ratios include choices related to negative book equity firms, price filters, and utility firms. Including financial firms increases the Sharpe ratio, on average, from 0.62 to 0.66. It can further be seen that eliminating industry exposures from factor returns substantially increases Sharpe ratios, which is in line with Daniel et al. (2020).

Equal-weighting portfolios improves the Sharpe ratio compared to value-weighting portfolios from 0.57 to 0.71, on average, and also comes out as one of the more important design choices. The Sharpe ratios for independent and dependent sorts are approximately similar. Finally, using the most recent market capitalization to construct factors increases the Sharpe ratio from 0.63 to 0.65 relative to using the market cap in June. Overall, our findings imply that construction choices can materially affect factor performance, especially those concerning NYSE breakpoints, micro stocks, industry-adjusted characteristics, and value-weighting.

4.2 Non-standard errors versus standard errors

Based on the above analyses, non-standard errors might 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. (2021)).

We initially 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 Figure 4. The white bars indicate the non-standard error 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 6 out of 11 factors, we find that the non-standard error is larger than the standard error. These factors are HML, HML(m), CMA, IA, ROE 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 0.04 and 0.09). In terms of proportions (non-standard error divided by average standard error), we find that this proportion ranges between 58% (SMB) and 190% (PEAD), with the average being 118%. Overall, we conclude that non-standard errors are sizable in comparison with standard errors. The average non-standard error to standard error ratio of 118% is also relevant in comparison to the ratio of 160% found by Menkveld et al. (2021), based on a relatively high degree of researcher discretion in their experiment.

The above estimation of non-standard errors hinges on the assumption that each design choice is made with an equal probability. Consequently, each possible combination occurs once out of 2048 times in our sample. From Table 2 we know that various choices do not occur with an equal probability in the literature. We can use these observed probabilities in estimating an alternative non-standard error, which we call the weighted non-standard error. More precisely, we use the implied probabilities to compute the total probability that an outcome of choices would occur. This total probability is computed by multiplying the individual implied probabilities. Subsequently, we compute the non-standard error as the weighted standard deviation across all construction choices, whereby we weight the observation by its total implied probability. We also use this weighting when calculating standard errors. We plot the weighted standard and non-standard errors in Figure 5 for each factor.

We find that the typical non-standard error remains sizable. The weighted non-standard errors exceed the estimated average weighted standard error in 5 out of 11 factors. Compared to Figure 4, the non-standard error of the ROE factor now falls below the standard error for that factor. The non-standard errors of the HML and PEAD factor are also slightly reduced compared to the non-weighted analysis. Overall, though, the results shown in Figure 5 and Figure 4 are very similar. The average weighted non-standard error divided by the average weighted standard error now ranges between 62% (FIN) and 189% (CMA), with the average being 108%.

4.3 Multiple hypothesis testing

Harvey et al. (2016) argue that as many studies attempt to explain the cross-section of expected returns, statistical inference should not be based on a "single" test perspective. For a single hypothesis test (one factor), a significance level (α) is used to control the type I error rate, i.e., the probability of finding a significant factor when it actually has no explanatory power. When there are many factors to be tested, it is likely for an event with probability α to occur. Hence, a type I error rate at α for an individual test is not sufficient to control for the overall probability of false discoveries. Multiple hypothesis testing takes into account that the family-wise (joint) error rate must be adjusted to simultaneously evaluate the outcomes of many individual tests.

To account for multiple hypothesis testing, we implement a simple Bonferroni adjustment. Under the Bonferroni adjustment, the significance level is divided by the number of hypothesis. Hence, we reject any hypothesis with a p-value smaller or equal to α/M. This adjustment applies the same adjustment to each test, whereby we inflate the original p-value by the number of tests M. In addition, we apply the Holm (1979) adjustment. Holm's adjustment is a step-down procedure, which is less stringent than Bonferroni. Since less stringent hurdles are applied, more discoveries are generated under Holm's than Bonferroni's adjustment.

The dimensionality of tests increases considerably with the amount of construction choices that are available: there are M factors, which each can be constructed in N possible ways. Hence, a researcher is testing $M \times N$ hypotheses, thereby increasing the hurdle to reject a hypothesis. A researcher should take this into account when testing multiple hypotheses. In our setting, we have 11 factors and 2048 construction sets, resulting in 22528 hypotheses to be tested. To take into account that factors can be constructed in many different ways, in this section we implement simple corrections to the significance levels.

The results are reported in Table 5. We test, for each factor, whether the Sharpe ratio is significantly different from zero. We report how many times a factor rejects the null hypothesis of not being different from zero. Under the classical hypothesis testing ("CHT"), we would have used a critical t-statistic of 1.96. Under this criterion, most factors typically reject the null hypothesis of a zero Sharpe ratio, although the SMB factor only rejects the null hypothesis in 9% of the cases. All other factors have rejection rates above 65%, with 7 of the 11 factors having rejection rates above 95%. When using the Bonferroni correction within a factor, we first employ a significance level of $\alpha/2048$, representing 2048 construction sets. This correction leads to a critical t-statistic of 4.25. The SMB factor is never statistically different from zero under this Bonferroni correction. The null hypothesis is also less often rejected for the other factors. For example, the HML factor rejects the null hypothesis with a probability of 78% under the CHT, but with a 16% probability under the Bonferroni correction. Next, we impose that $M \times N$ hypotheses are tested. Using a Bonferroni correction leads to a significance level of $\alpha/22528$, which corresponds to a critical t-value of 4.78. Not surprisingly, rejection rates of the null hypothesis decrease further. For example, the null hypothesis for the UMD factor is now rejected in only 6% of the cases. As indicated by the two columns on the right-hand side of Table 5, our findings are similar when applying the Holm corrections.

5 Model Selection

The prior section has shown that Sharpe ratios within factors depend on a range of construction choices and that the non-standard errors surrounding portfolio sorting can be substantial. In this section, we study the implications of non-standard errors 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. (2021).

5.1 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:

$$Sh^2(f,\Omega) - Sh^2(f) = \alpha' \Sigma^{-1} \alpha \tag{3}$$

$Sh^2(f,\Omega)$ 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 Equation 3. 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. 3 corresponds to maximizing $Sh^2(f)$. Hence, model selection can be examined by comparing the maximum squared Sharpe ratio across models.[9]

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. Figure 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. We separate the value-weighted portfolio returns (dashed bars) and equal-weighted portfolio returns (white bars).

The average maximum Sharpe ratio for the mean-variance optimal FF5 model, using value-weighted returns, is 1.08. Replacing operating profitability with cash profitability

[9] Detzel et al. (2021) 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. Hence, we also consider net factor returns. We report all corresponding net return analyses in Appendix C.

increases this value to 1.35. 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, 1.67, and 1.71, respectively. Based on these averages, the preferred model would be the DHS factor model, with the BS6 model coming very close. When factors are equally weighted, factor models have higher maximum Sharpe ratios, on average. The highest average maximum Sharpe ratios with equal-weighted returns are also obtained by the BS6 and DHS factor models, both with values of roughly 2, and now the BS6 model has a slightly higher average value.

Differences in construction choices induce non-standard errors in factor premiums and subsequently also 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 equally-weighted BS6 model, we find a 95% confidence interval between 1.57 and 2.54. Due to the non-standard errors, model rankings may differ across different sets of construction choices. We find that in 39.9% of all choice sets, the BS6 has the largest maximum Sharpe ratio. The DHS model has the largest maximum Sharpe ratio in 60.1% 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 differential choices across models (for example, 80-20 breakpoints for the PEAD factor but 70-30 breakpoints for factors in another model). In Section 5.4, we use a bootstrap approach to further study how often one model outperforms the others.

Table 6 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 47.6% 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 27.6% and 77.6% for a two standard deviation change. HML has a small average weight of 1.5% in the 5-factor model. Interestingly, for some construction methods the HML weight is negative (-18.1% for a two standard deviation decrease) while for others it is substantially positive (21.1% 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 4%. For example, the I/A factor ranges between 31.7% and 46.9%, given a two standard deviation interval. The BS6 model aims to improve on the Fama-French models by adding the monthly updated value factor, which correlates more negatively with momentum. Consequently, the UMD factor receives a larger average weight of 21.1%, with a standard deviation of 4.1% across construction methods. The monthly updated value factor receives a relatively larger weight (compared to FF-models) of 26.7%, with a standard deviation of 7.3%. On average, the PEAD factor is important in the DHS model. The model allocates on average 58.1% to the PEAD factor, with a standard deviation of 6.5%.

5.2 Efficient frontier expansion

The results from the previous subsection 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:

$$MVP_{M1 \cup M0,t} = \alpha + \beta MVP_{M0,t} + \epsilon_t \tag{4}$$

Table 7 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 0.13% and 0.45% per month, with t-statistics above 4.5. We especially find that adding the BS6 factors or DHS3 factors (M1) to FF models (M0) greatly improves the efficient frontier, with alphas between 0.25% and 0.45% 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 0.06% per month. However, the estimated alpha has a standard deviation of 0.05%. 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.

5.3 Economic significance

Next, we quantify the economic significance, in Table 8, 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 combined model. We find that the FF5 model can be improved between 20.5% up to 82.3%, 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 37.1% and 82.3%, whereas adding the DHS3 model yields gains between 36.2% and 74.7%. 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 26.0%, whereas vice versa the gain is 26.9%.

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 82.3%, but also has a substantial standard deviation of 17.0%. This implies that there is a construction set for which the improvement is 48.3%, but also a set for which the improvement is 116.3%. The Q-factor model improves the $FF6_c$ by 6.7% on average, with a standard deviation of 5.8%. 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 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.

5.4 In-sample and out-of-sample estimation

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 600 months into 300 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 300 pairs. We randomly assign a month from each pair to the IS sample.[10] 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

[10] Note that a month might appear multiple times in the IS sample if the pair is drawn multiple times.

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 9 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 300 observation months. We find that the $FF6_c$ model outperforms the Q factor model in 64.8% of the sample. Its Sharpe ratio (1.67) is slightly higher than that of the Q factor model (1.60). The BS6 model seems to outperform the other models, with pairwise win-probabilities over 50.7% and an average Sharpe ratio 0f 1.98. It is, on average, the model with the highest Sharpe ratio in 47.6% of all simulation runs. The standard deviation is 29.1%, implying large variation across construction methods. The DHS model in this simulation has an average Sharpe ratio of 1.92, making it the second-best model in this aspect. Still, this model has the largest Sharpe ratio in 47.8% of all simulation runs.

Panel B presents the out-of-sample (OS) results. We find that the DHS model outperforms all other models in 57.9% of the simulation runs in an out-of-sample setting, averaged across construction methods. The BS6 model obtains an overall win-probability of 38.4%, making it the second strongest model from an out-of-sample perspective. Both models have an almost 30% standard deviation 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.

6 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):

$$E(R_{long} - R_{short}) = \beta(F_{long} - F_{short})$$

where F stands for the factor characteristics of the long and short portfolio. 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:[11]

$$Normalized\ factor\ score_{i,v,t} = \frac{Variable_{i,v,t} - Mean_{v,t}}{Max_{v,t} - Min_{v,t}} \tag{5}$$

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 weighting scheme, value- or equal-weighted. 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 to an impact on factor exposure, construction choices may impact the liquidity

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

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:

$$ILLIQ_{i,t} = \frac{1}{D_{i,t}} \sum_{t=1}^{D_{i,t}} \frac{|R_{i,t,d}|}{VOLD_{i,t,d}} \tag{6}$$

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 their daily price series. We provide more details on this procedure in Appendix C.2. 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 10 shows the estimated coefficients.

Overall, most construction choices significantly impact portfolio characteristics. We find that 7, 7 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, 6 basis points lower transaction costs than 20-80 breakpoints portfolios. Using NYSE instead of NAN breakpoints significantly lowers transaction costs by an average of 15 basis points and improves portfolio liquidity. This is sensible 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 negatively 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. Value-weighting as opposed to equal-weighting significantly reduces factor exposure. This reduction is compensated by a significantly higher liquidity profile and significantly lower transaction costs.

7 Reducing Non-Standard Errors

The sizable non-standard errors documented in this paper represent considerable uncertainty created by the evidence-generating process related to portfolio sorting in asset pricing. Our analysis can reduce this uncertainty by providing guidelines on how to reduce non-standard errors. Minimizing non-standard errors would imply that the research field reaches a consensus on all research design choices, leading to researchers adopting a common set of procedures. However, there is a trade-off to be made in reaching such a consensus. Variation in 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 type of firm. Still, allowing for too many degrees of freedom regarding design choices and the resultant high non-standard errors could induce excessive reporting of statistically significant results (i.e., p-hacking).

In this section, we take this tradeoff into account and examine whether a limited set of restrictions could substantially reduce non-standard errors. This analysis follows from Figure 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 Section 3.2. In “Set 2” we exclude three choices that appear particularly important in Figure 3: NAN breakpoints, including microcaps, and equal-weighting. Excluding these three choices could substantially reduce uncertainty in interpreting reported results and the choices can also be justified relatively easily based on economic arguments. For instance, set 2 resembles the choices made by Hou et al. (2020). 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 segment of the market difficult for investors. The other choices link to microcaps. When researchers opt for equal-weighting portfolios, microcaps (and small caps) become relatively important, which tends to bias the mean return upward. This bias is limited when value-weighted returns are computed. Using NAN breakpoints also favors micro- and small caps, leading to similarly inflated anomaly profits. Excluding these three choices leaves researchers with eight remaining design choices, or 256 possible combinations.

A fourth choice that seems important for non-standard errors is industry neutralization. However, here the trade-off is especially important. Figure 3 suggests that the conservative choice is to not use industry-adjusted characteristics. However, Daniel, Mota, et al. (2020) suggest that this tends to pick-up unintended (industry) risks, generating portfolios that are no longer mean-variance efficient. Hedging this exposures is a choice that can be made in order to improve risk-adjusted returns. However, this choice could depend on the particular research question one is after, and the unhedged approach is the more popular approach, as shown by the results in Table 2.

Instead, 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)). Fama and French (1992) have also made it common practice to construct size-breakpoints based on the market capitalization of firms at the end of June. Set 3 thus only leaves four choices open: whether to impose price filters (but this 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.

Figure 7 shows the computed non-weighted non-standard errors per factor for set 1 (the base case), set 2 and set 3. For each factor, we find that the non-standard error can be heavily reduced by imposing the restrictions of set 2, i.e., the use of NYSE breakpoints, excluding microcaps, and using value-weighting. For example, the PEAD factor has a non-standard error of above 0.10 using the original set of eleven choices, which decreases to about 0.02 (a 76% decline) for set 2. On average, across factors, we find that non-standard errors decrease by 70% when moving from set 1 to set 2. Set 3 does not yield a substantial additional decline in non-standard errors for most factors. For some factors, the non-standard errors are even higher for set 3 than for set 2. On average, set 3 leads to a 73% 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 employ value-weighting. 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 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.

8 Conclusion

Within empirical asset pricing, character-based sorting is a popular way to construct 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 find that construction choices indeed impact factor returns. Using 2048 different combinations of construction 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 adds a layer of uncertainty in academic work, leading to non-standard errors. 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. An alternative calculation of non-standard errors that takes the popularity of choices into account reinforces this conclusion.

The variation that we document materially impacts model selection exercises when comparing models. Mean-variance weights vary substantially across construction methods, with some factors receiving zero weight when taking transaction costs into account. Maximum Sharpe ratios of factor models also show wide variation across construction methods. By following a bootstrapping approach, we show that the probability of having the highest Sharpe ratio has standard deviations of up to 41%. In addition, our study points out other important consequences of using specific construction choices, such as those related to factor and liquidity exposure, and transaction costs.

Our study has important implications for researchers. As in Mitton (2022), who focuses on methodological variation in empirical corporate finance, robustness tests are important, especially if researcher discretion on which robustness results to report is limited. Our results suggest that the most important design choice around factor construction are those concerning NYSE or NAN breakpoints, micro stocks, industry-adjusted characteristics, and value-weighting. In a specification check (Brodeur et al. (2020)), researchers could graphically show the distribution of their Sharpe ratios (or other results) if their design choices are varied among these four dimensions. Alternatively, results become easier to compare when researchers make similar choices. Based on our analysis, we recommend the consistent use of NYSE breakpoints, exclusion of microcaps, and value-weighting. Following this guideline reduces the average non-standard error by 70%. Our analysis further 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 winning model depends on the construction choices being made. In short, our main recommendation is that researchers should mind their sorts.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

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

Asness, C. S., Porter, R. B., & Stevens, R. L. (2000). Predicting stock returns using industry-relative firm characteristics. *Available at SSRN 213872*.

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

Barillas, F., Kan, R., Robotti, C., & Shanken, J. (2020). Model comparison with sharpe ratios. *Journal of Financial and Quantitative Analysis*, *55*(6), 1840–1874.

Barillas, F., & Shanken, J. (2017). Which alpha? *The Review of Financial Studies*, *30*(4), 1316–1338.

Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance*, *73*(2), 715–754.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634–3660.

Brown, S. J., Lajbcygier, P., & Li, B. (2008). Going negative: What to do with negative book equity stocks. *Journal of Portfolio Management*, *35*(1), 95–102.

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

Daniel, K., Hirshleifer, D., & Sun, L. (2020). Short-and long-horizon behavioral factors. *The Review of Financial Studies*, *33*(4), 1673–1736.

Daniel, K., Mota, L., Rottke, S., & Santos, T. (2020). The cross-section of risk and returns.

The Review of Financial Studies, *33*(5), 1927–1979.

Detzel, A. L., Novy-Marx, R., & Velikov, M. (2021). Model selection with transaction costs. *Available at SSRN 3805379*.

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 3959116*.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, *63*(4), 1653–1678.

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

Fama, E. F., & French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, *128*(2), 234–252.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, 1121–1152.

Giglio, S., Liao, Y., & Xiu, D. (2021). Thousands of alpha tests. *The Review of Financial Studies*, *34*(7), 3456–3496.

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

Harvey, C. R., & Liu, Y. (2019). A census of the factor zoo. *Available at SSRN 3341728*.

Harvey, C. R., & Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, *75*(5), 2503–2553.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *The Journal of Finance*, *64*(3), 1445–1477.

Hirshleifer, D., & Jiang, D. (2010). A financing-based misvaluation factor and the cros-section of expected returns. *Review of Financial Studies*, *23*(1), 3401–3436.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Hou, K., Mo, H., Xue, C., & Zhang, L. (2019). Which factors? *Review of Finance*, *23*(1), 1–35.

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650–705.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, *33*(5), 2019–2133.

Kessler, S., Scherer, B., & Harries, J. P. (2020). Value by design? *Journal of Portfolio Management*, *46*(2), 1–19.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292.

Lee, C., & Swaminathan, B. (2000). Price momentum and trading volume. *Journal of Finance*, *55*(1), 2017–2069.

Li, J., & Robotti, C. (2022). On the power of asset pricing tests. *Working paper*.

Linnainmaa, J. T., & Roberts, M. R. (2018). The history of the cross-section of stock returns. *The Review of Financial Studies*, *31*(7), 2606–2649.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5–32.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., . . . Weitzel, U. (2021). Non-standard errors. *Working paper*.

Mitton, T. (2022). Methodological variation in empirical corporate finance. *The Review of Financial Studies*, *35*, 527–575.

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

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, *29*(1), 104–147.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, *39*(4), 1127–1139.

Stambaugh, R. F., & Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, *30*(4), 1270–1315.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics*, *99*(1), 1–10.

Wahal, S., & Yavuz, M. D. (2013). Style investing, comovement and return predictability. *Journal of Financial Economics*, *107*(1), 136–154.

Yan, X. S., & Zheng, L. (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *The Review of Financial Studies*, *30*(4), 1382–1423.

Figure 1: **Construction choices and Sharpe ratios of the HML factor.** This figure plots annualized gross Sharpe ratios for long-short factor returns, where a factor is constructed by using 2048 different factor construction methods. The red dot shows the median Sharpe ratio for the HML factor in our sample. The blue dot shows the Sharpe ratio using the construction choices mentioned in the original study. The sample runs from January 1972 until December 2021.

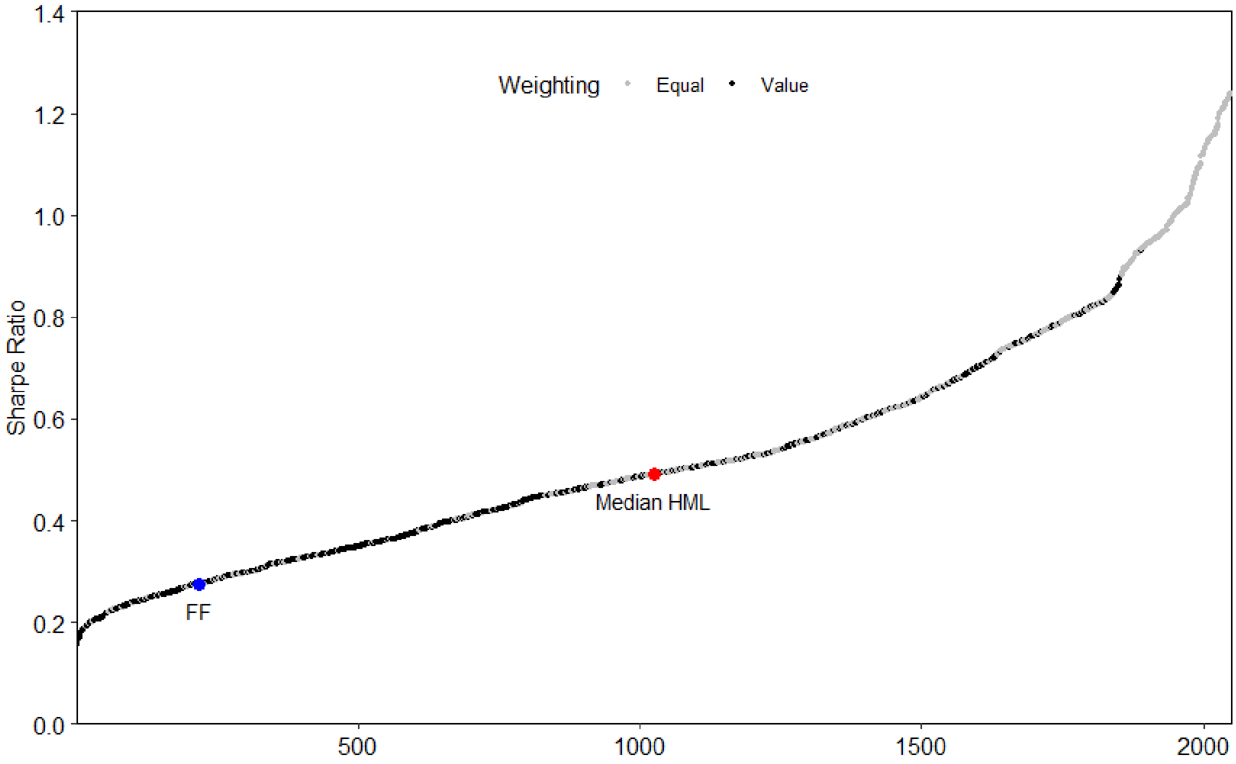
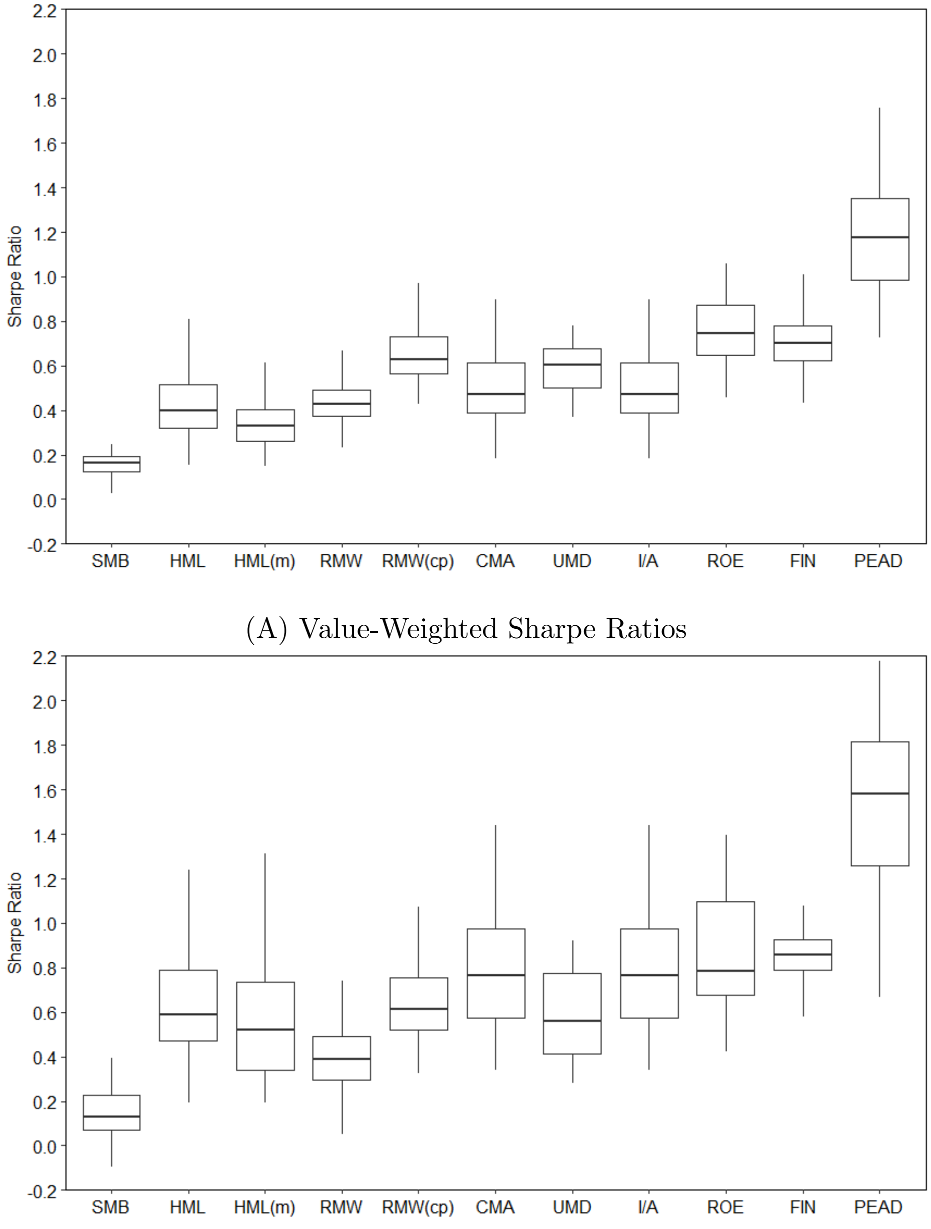


Figure 2: **Sharpe ratio variation within factors.** This figure plots the distribution of annualized value-weighted (subfigure A) and equal-weighted (subfigure A) gross Sharpe ratios for long-short factor returns, where a factor is constructed 2048 times by using the 2048 different factor construction 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 January 1972 until December 2021.



(A) Value-Weighted Sharpe Ratios

(B) Equal-Weighted Sharpe Ratios

Figure 3: **Construction choices and Sharpe ratios.** This figure shows the impact of construction choices on the gross Sharpe ratio averaged over factors. Sharpe ratios are annualized. We consider eleven choices. "30-70" refers to the use of the 30th and 70th percentile as breakpoints in the sorting procedure ("Yes") or the use of the 20th and 80th percentile ("No"). NYSE indicates whether the NYSE cross-section is used to construct breakpoints ("Yes") or the full NYSE-AMEX-Nasdaq cross-section ("No"). "BE" indicates whether stocks with negative book equity are excluded ("Yes") or included ("No"). "Micro" indicates whether we include stocks with the smallest 20% market capitalization ("Yes") or not ("No"). "PRC" indicates whether stocks with a price below 5 dollar are excluded ("Yes") or included ("No"). "Utilities" means that companies in the utility sector are included ("Yes") or excluded ("No"). "Financial" means that companies in the finance sector are included ("Yes") or excluded ("No"). "Ind_Neutral" means that portfolio sorts are constructed using industry-adjusted characteristics ("Yes") or the standard characteristics ("No"). "VW" indicates whether factors are calculated by using value-weighting ("Yes") or equal-weighting ("No"). "Independent" refers to the use of independent sorting ("Yes") or dependent sorting ("No"). "Recent" indicates that we use the one-month lagged market capitalization ("Yes") or the market capitalization of June ("No"). Monthly factor returns are constructed using data from January 1972 to December 2021.

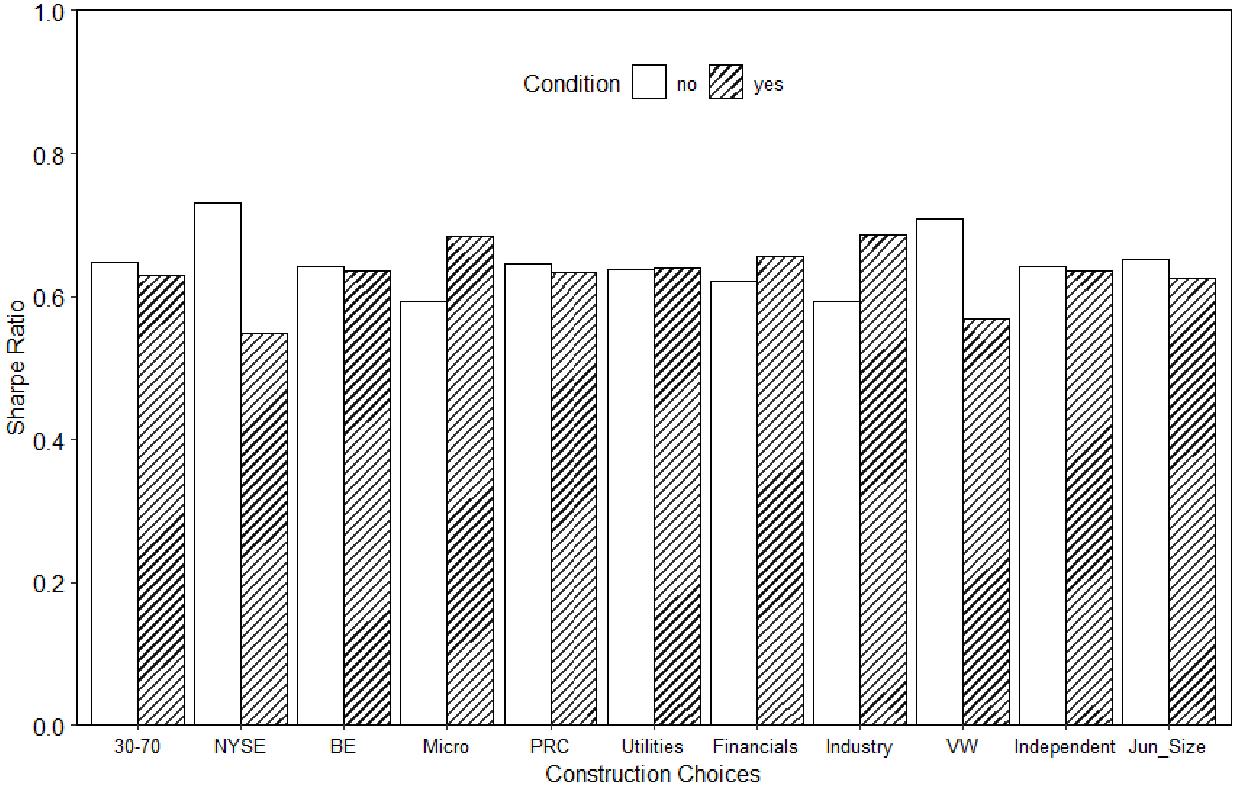


Figure 4: **Non-standard errors and standard errors.** This figure plots the non-standard error (white) and standard error (dashed bar) for each factor. The non-standard error is defined as the cross-sectional standard deviation of Sharpe ratios, where the cross-section consist of all 2048 versions of 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 January 1972 and December 2021.

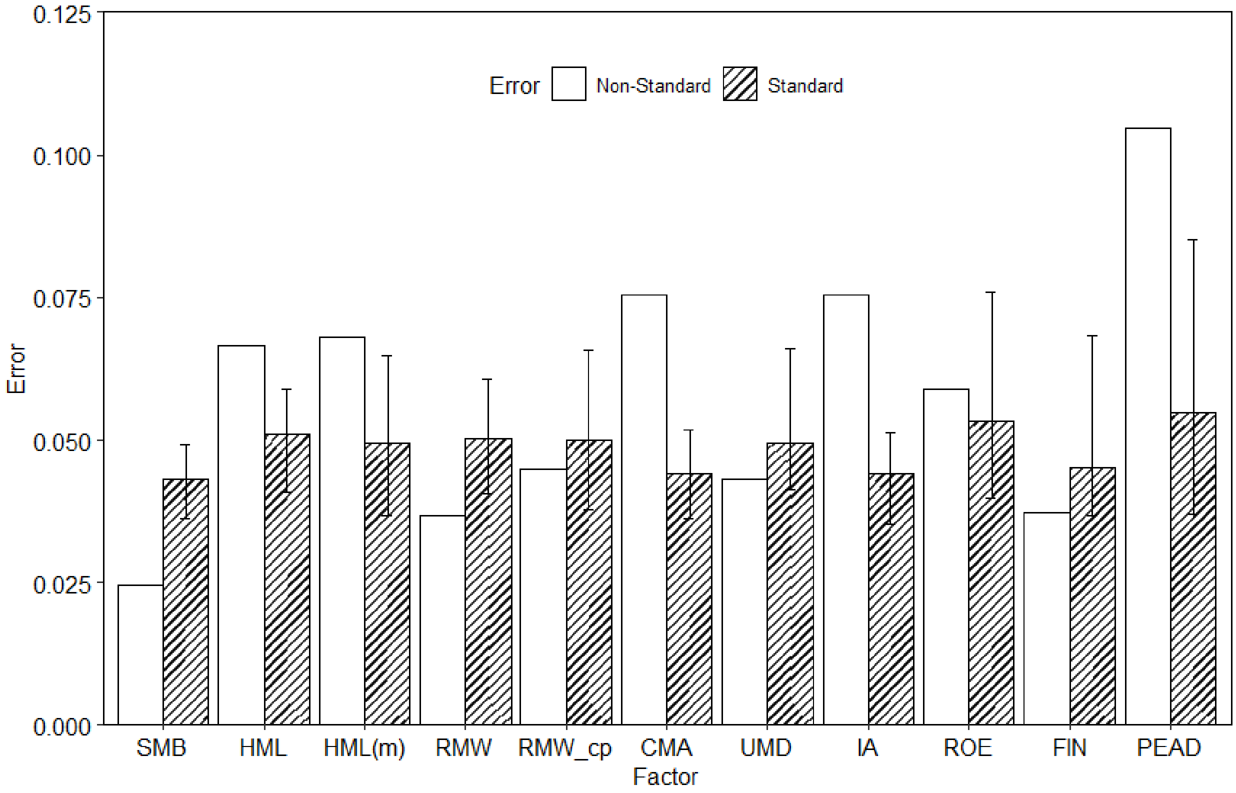


Figure 5: **Weighted non-standard errors and standard errors.** This figure plots the non-standard error (white) and standard error (dashed bar) for each factor. The non-standard error is defined as the cross-sectional weighted standard deviation of Sharpe ratios, where the cross-section consist of all 2048 versions of a factor. The standard error is the weighted 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. We weight the errors by the survey-implied probabilities. 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 January 1972 and December 2021.

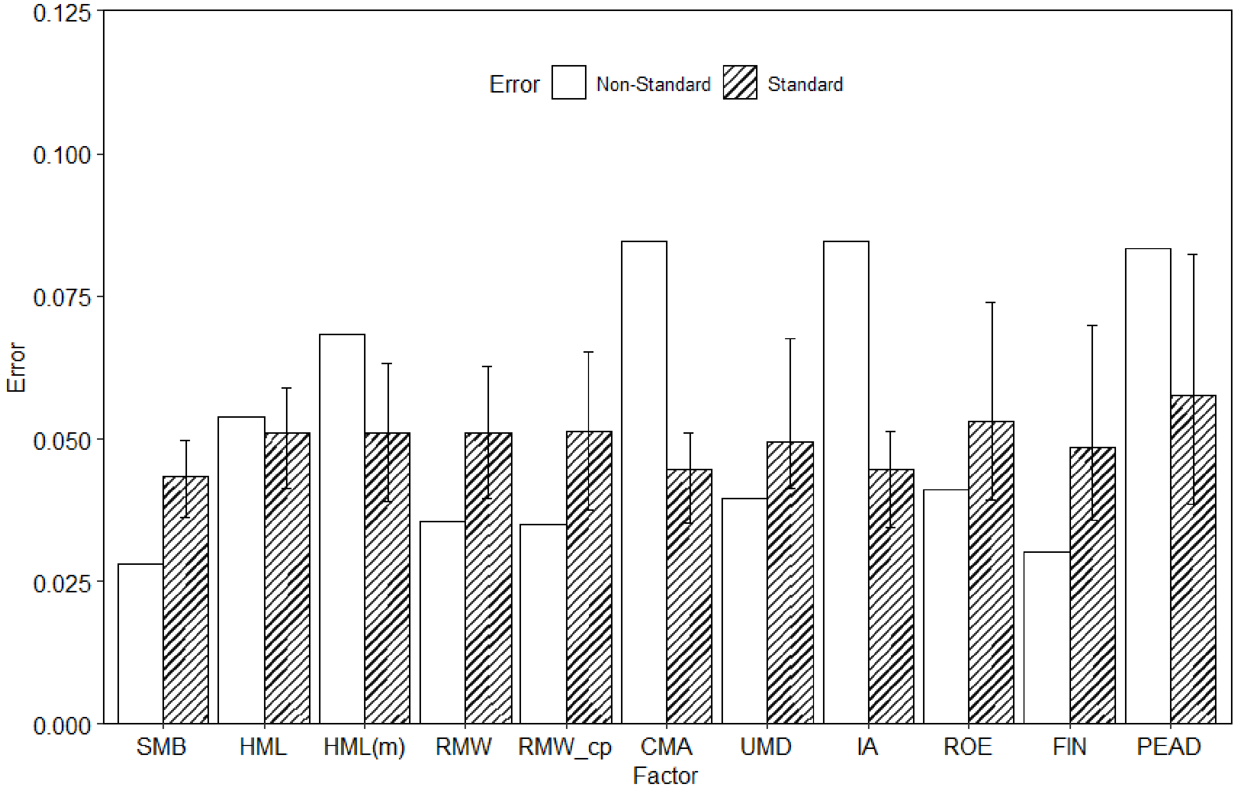


Figure 6: **Selecting factor models.** This figure shows the maximum gross Sharpe ratio (annualized) from the factors from the factor models listed on the horizontal axis. The white bar shows the maximum Sharpe ratio obtained by using equal weighted factor returns. The dashed bar shows the maximum Sharpe ratio using value weighted factor returns. The error plot shows the variation in the maximum Sharpe ratios for a given factor model, across construction choices. The data runs from January 1972 until December 2021.

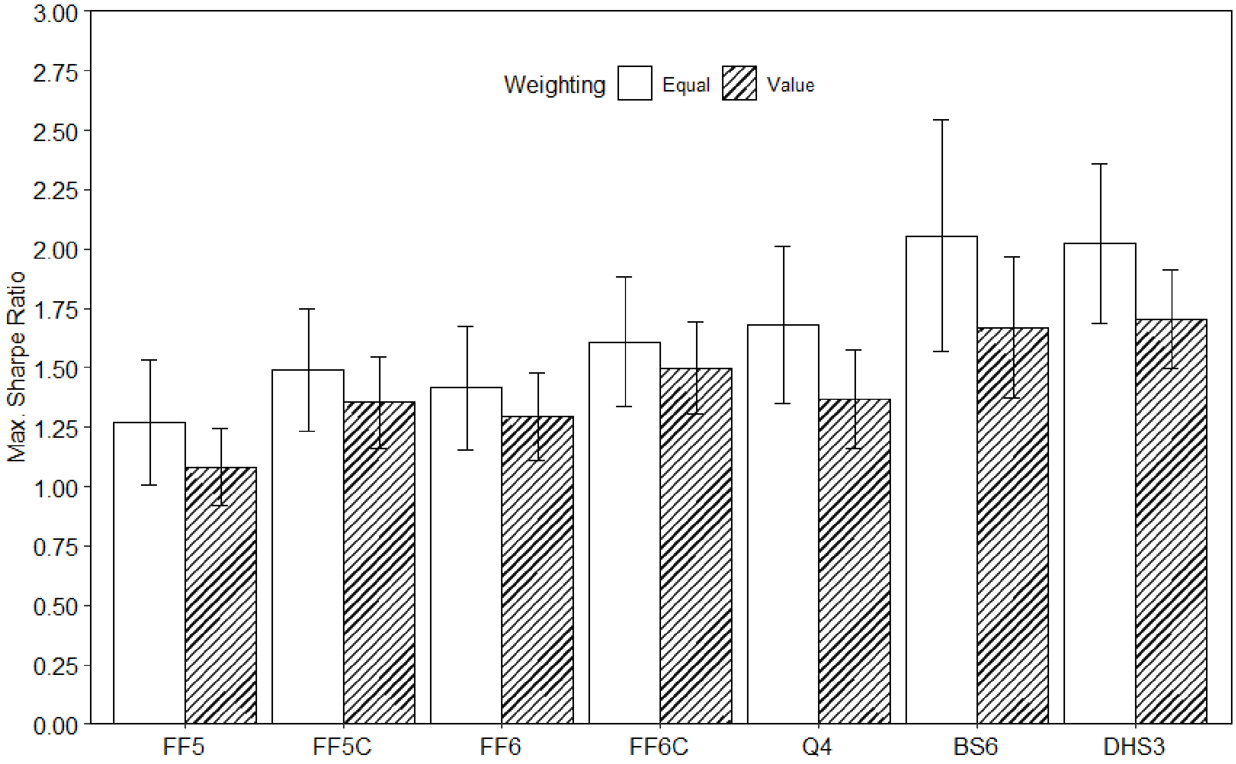


Figure 7: **Reducing non-standard errors.** The figure shows the non-standard errors using three sets of research design choices. Set (1) includes all eleven construction choices. Set (2) imposes NYSE breakpoints, excludes micro-caps, and imposes value-weighting. Set (3) extends on set (2) by further imposing the use of 30-70 breakpoints, the exclusion of firms with a negative book value and financial firms, and by imposing the measurement of size in June.

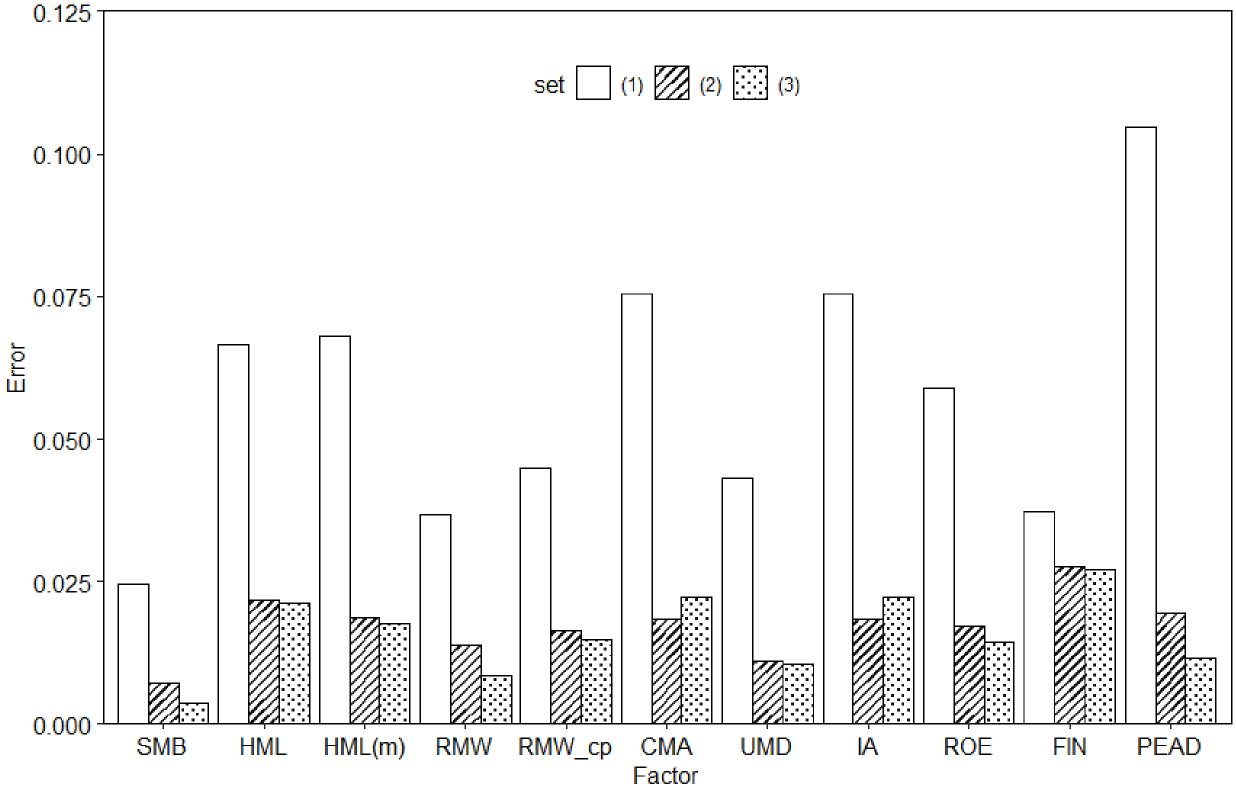


Table 1: **Factor models.** This table lists the non-market factors used by asset pricing models, indicated by a ✓. It also lists properties of the factor construction methodology: the sorting characteristic, the breakpoints (BP), the rebalancing frequency (Rebalancing), and the sorting method (Construction). In each model, factor returns are defined as the equal-weighted average of the returns on the portfolios with high (or low) values of the primary sorting characteristic minus the equal-weighted average of the portfolios with low (or high) values. SMB returns are given by the simple average of the returns on all portfolios with low size minus the average of the returns on all portfolios with large size in three independent 2x3 sorts of stocks on size and each of the following characteristics: book-to-market, growth in book assets, and operating profitability. ME returns are given by the simple average of the returns on all portfolios with low size minus those with large size in 2x3x3 sorts on size, growth in book assets, and return on equity. 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 q-model. BS6 denotes the Barillas and Shanken (2018) six-factor model. DHS3 denotes the Daniel, Hirshleifer, and Sun (2020) three-factor model.

					Factor models						
Factor	Sorting characteristic	BP	Rebalancing	Construction	FF5	FF6	FF5$_c$	FF6$_c$	Q4	BS6	DHS3
SMB	Market capitalization	50-50	Annual	2x3	✓	✓	✓	✓	✓	✓	
HML	Book-to-market	70-30	Annual	2x3	✓	✓	✓	✓			
HML(m)	Book-to-market	70-30	Monthly	2x3						✓	
RMW	Accruals operating profitability	70-30	Annual	2x3	✓	✓					
RMW(cp)	Cash operating profitability	70-30	Annual	2x3			✓	✓			
CMA	Growth in book assets	70-30	Annual	2x3	✓	✓	✓	✓			
UMD	$R_{t-12,t-2}$	70-30	Monthly	2x3		✓		✓		✓	
I/A	Growth in book assets	70-30	Monthly	2x3x3					✓	✓	
ROE	Quarterly returns-on-equity	70-30	Monthly	2x3x3					✓	✓	
FIN	Net and composite share issuance	80-20	Annual	2x3							✓
PEAD	4-day CAR earnings announcements	80-20	Monthly	2x3							✓

Table 2: **Variation in empirical finance.** This table shows the results from surveying methodological choices that have been made in the empirical asset pricing literature. We report the proportions of the choice (1) or (2) occuring in 323 empirical articles in the top finance journals between 1965 and 2018, based on the list of papers that Campbell Harvey and Yan Liu constructed for their census of the factor zoo (Harvey and Liu (2019)).

	Options		**Proportion**	
	(1)	(2)	(1)	(2)
Choice 1	Use 30-70 BP	Use 20-80 BP	49.3%	50.7%
Choice 2	Use NYSE BP	Use NAN BP	41.5%	58.5%
Choice 3	Exclude $BE < 0$	Include $BE < 0$	22.0%	78.0%
Choice 4	Include Microcaps	Exclude Microcaps	88.2%	11.8%
Choice 5	Impose price filter	No price filter	18.3%	81.7%
Choice 6	Include utilities	Exclude utilities	90.1%	9.9%
Choice 7	Include financials	Exclude financials	71.2%	28.8%
Choice 8	Industry Neutrality	Unhedged	11.5%	88.5%
Choice 9	Value-Weighted	Equal-Weighted	58.5%	41.5%
Choice 10	Independent	Dependent	71.8%	28.2%
Choice 11	June Size	Recent Size	67.4%	32.6%

Table 3: **Correlation of choices.** This table reports the correlation matrix between the eleven methodological choices in our sample.

Choice	30-70	NYSE	BE	Micro	PRC	Utilities	Financials	Industry	VW	Independent	Jun_Size
30-70	1										
NYSE	0.12	1									
BE	-0.05	0.32	1								
Micro	0.02	0	0.02	1							
PRC	0.07	-0.14	-0.19	-0.42	1						
Utilities	-0.08	0.18	0.15	-0.1	0.03	1					
Financials	-0.17	-0.30	-0.53	-0.14	0.25	0.32	1				
Industry	0.06	-0.18	-0.12	-0.19	0.21	-0.05	0.12	1			
VW	0.12	0.19	0.24	0.22	-0.15	-0.17	-0.3	-0.12	1		
Independent	0.30	0.14	0.26	-0.02	-0.06	0.11	-0.31	0.03	0.14	1	
Jun_Size	0.28	0.27	0.38	0.19	-0.23	0.07	-0.43	-0.09	0.31	0.54	1

Table 4: **Summary statistics.** This table reports the annualized average return (in %) and Sharpe ratio of the factors listed in Table 1, gross of transaction costs. We report these statistics for both the value-weighted and equal-weighted models. The data runs from January 1972 until December 2021.

	Value-Weighted		**Equal-Weighted**	
	Mean	Sharpe	Mean	Sharpe
SMB	1.91	0.18	0.86	0.11
HML	3.37	0.36	4.85	0.53
HML(m)	3.37	0.30	4.66	0.42
RMW	3.84	0.47	4.13	0.50
RMW(cp)	4.66	0.70	5.14	0.77
CMA	3.09	0.46	3.96	0.64
UMD	8.00	0.59	9.08	0.70
IA	3.09	0.46	3.96	0.64
ROE	7.43	0.80	9.14	1.03
FIN	7.51	0.72	8.90	0.89
PEAD	5.96	1.10	7.01	1.55

Table 5: **Multiple hypothesis testing and empirical rejection rates.** This table reports the empirical rejection rates using both classical hypothesis testing and multiple hypothesis testing. For each factor, we report its Sharpe ratio averaged over our construction sets ("Sharpe") and probability that the Sharpe ratio is smaller or equal than zero ($Pr(Sh <= 0)$). We block-bootstrap all series 10.000 times to calculate the statistics. "CHT" reports the proportion of Sharpe ratios for which the null hypothesis of a zero Sharpe ratio is rejected, using a significance level of 5%. $Bonf_w$ and $Bonf_a$ reports this proportion when correcting α to $\alpha/2048$ and $\alpha/(2048 * 11)$, respectively. $Holm_w$ and $Holm_a$ applies the Holm correction using 2048 and 2048*11 hypotheses, respectively. The data runs from January 1972 until December 2021.

Factor	Sharpe	Pr($Sh <= 0$)	CHT	Bonf$_w$	Bonf$_a$	Holm$_w$	Holm$_a$
SMB	0.05	17.99	0.09	0.00	0.00	0.00	0.00
HML	0.15	1.71	0.78	0.16	0.10	0.17	0.11
HML(m)	0.14	2.41	0.65	0.14	0.09	0.15	0.09
CMA	0.18	0.37	0.96	0.42	0.30	0.44	0.31
RMW	0.12	3.22	0.73	0.00	0.00	0.00	0.00
RMW(cp)	0.19	0.13	0.99	0.25	0.14	0.28	0.15
UMD	0.17	0.29	0.97	0.21	0.06	0.23	0.08
IA	0.18	0.36	0.96	0.41	0.30	0.45	0.31
ROE	0.23	0.01	1.00	0.48	0.30	0.59	0.32
FIN	0.22	0.00	1.00	0.77	0.58	0.93	0.61
PEAD	0.39	0.00	1.00	0.97	0.96	1.00	0.96

Table 6: **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 January 1972 to December 2021.

	Mkt	SMB	HML	RMW	CMA	UMD	RMW_{cp}	IA	ROE	HML_d	FIN	PEAD
FF5	20.2	5.6	1.5	25.1	47.6							
	(3.9)	(3.0)	(9.8)	(6.4)	(10.0)							
FF6	17.9	6.2	9.7	20.6	29.3	16.3						
	(3.6)	(3.2)	(8.8)	(6.2)	(10.2)	(3.8)						
FF5C	18.1	9.4	-0.5		33.3		39.7					
	(3.1)	(2.8)	(9.3)		(10.7)		(7.4)					
FF6C	16.8	9.1	6.4		21.5	13.0	33.2					
	(3.0)	(2.7)	(8.7)		(9.9)	(3.2)	(6.6)					
Q4	17.3	11.2						39.3	32.1			
	(2.5)	(2.0)						(3.8)	(3.8)			
BS6	14.4	7.2				21.1		9.6	21.0	26.7		
	(3.3)	(3.0)				(4.1)		(7.0)	(4.5)	(7.3)		
DHS3	17.7										24.1	58.1
	(2.8)										(5.0)	(6.5)

Table 7: **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 January 1972 until December 2021.

	Union Model (M1)						
Base Model (M0)	FF5	FF5$_c$	FF6	FF6$_c$	Q4	BS6	DHS3
FF5	0.00	0.13	0.14	0.21	0.27	0.45	0.36
	(0.00)	(4.89)	(4.86)	(6.79)	(7.48)	(11.87)	(10.59)
	[0.00]	[0.04]	[0.05]	[0.06]	[0.11]	[0.15]	[0.09]
FF5$_C$	0.03	0.00	0.13	0.11	0.13	0.35	0.31
	(2.13)	(0.00)	(4.93)	(4.39)	(5.28)	(10.50)	(9.73)
	[0.03]	[0.00]	[0.05]	[0.04]	[0.07]	[0.13]	[0.09]
FF6	0.00	0.09	0.00	0.09	0.16	0.36	0.29
	(0.00)	(4.35)	(0.00)	(4.35)	(5.14)	(8.39)	(9.31)
	[0.00]	[0.03]	[0.00]	[0.03]	[0.12]	[0.16]	[0.09]
FF6$_c$	0.02	0.00	0.02	0.00	0.06	0.27	0.25
	(1.82)	(0.00)	(1.82)	(0.00)	(3.37)	(7.44)	(8.81)
	[0.02]	[0.00]	[0.02]	[0.00]	[0.06]	[0.14]	[0.09]
Q4	0.06	0.03	0.09	0.08	0.00	0.18	0.26
	(2.89)	(2.33)	(3.72)	(3.75)	(0.00)	(6.33)	(8.16)
	[0.05]	[0.03]	[0.06]	[0.04]	[0.00]	[0.08]	[0.07]
BS6	0.16	0.14	0.16	0.14	0.00	0.00	0.22
	(5.09)	(5.07)	(5.09)	(5.07)	(0.00)	(0.00)	(8.10)
	[0.11]	[0.08]	[0.11]	[0.08]	[0.00]	[0.00]	[0.06]
DHS	0.09	0.11	0.10	0.12	0.11	0.20	0.00
	(4.23)	(5.06)	(4.31)	(5.10)	(4.79)	(7.41)	(0.00)
	[0.08]	[0.07]	[0.08]	[0.07]	[0.07]	[0.11]	[0.00]

Table 8: **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 January 1972 until December 2021.

	Union Model (M1)						
Base Model (M0)	FF5	FF5_c	FF6	FF6_c	Q4	BS6	DHS3
FF5		20.5	21.8	36.8	36.6	82.3	74.7
		(9.0)	(8.4)	(12.2)	(13.1)	(17.0)	(15.5)
FF5_c	4.0		18.0	15.1	16.5	57.1	55.6
	(3.8)		(7.0)	(6.1)	(7.8)	(16.9)	(13.5)
FF6	0.0	12.2		12.2	16.0	50.9	45.0
	(0.0)	(4.9)		(4.9)	(10.6)	(21.6)	(10.1)
FF6_c	2.5	0.0	2.5		6.7	37.1	36.2
	(2.4)	(0.0)	(2.4)		(5.8)	(19.6)	(9.2)
Q4	5.4	4.1	8.7	9.8		22.0	37.3
	(4.1)	(3.8)	(4.2)	(6.5)		(10.5)	(8.5)
BS6	15.9	15.3	15.9	15.3	0.0		26.9
	(9.7)	(7.8)	(9.7)	(7.8)	(0.0)		(6.2)
DHS3	9.5	12.8	10.7	13.8	11.4	26.0	
	(7.7)	(7.6)	(8.7)	(8.4)	(7.1)	(16.2)	

Table 9: **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. We use the factor models listed in Table 1. “SR” reports the maximum Sharpe ratio of the row model, averaged across construction methodologies. $\sigma(SR)$ reports the standard deviation of the maximum 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 estimates and Panel B shows the out-of-sample estimates using gross returns. The estimates are based on 100.000 in-sample and out-of-sample simulation runs. Each simulation run splits the 600 sample months, running from January 1972 until December 2021, into 300 adjacent pair-months. The run randomly draws a sample of pairs (with replacement). The in-sample simulation randomly draws one month from each pair within a run. The remaining months form the out-of-sample. The in-sample observations are used to calculate in-sample Sharpe ratios and portfolio weights. The in-sample portfolio weights are applied to the out-of-sample returns to produce an out-of-sample 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	1.9	0.0	0.1	2.2	0.1	2.8	0.00	0.00	1.27	0.23
FF6	98.1	0.0	30.3	0.0	22.6	2.3	8.3	0.00	0.00	1.45	0.23
FF5$_c$	100.0	69.7	0.0	4.4	38.8	0.9	10.1	0.04	0.12	1.54	0.23
FF6$_c$	99.9	100.0	95.6	0.0	64.8	10.6	20.7	4.57	7.34	1.67	0.24
Q4	97.8	77.4	61.2	35.2	0.0	0.0	14.0	0.00	0.00	1.60	0.32
BS6	99.9	97.7	99.1	89.4	100.0	0.0	50.7	47.61	29.06	1.98	0.44
DHS	97.2	91.7	89.9	79.3	86.0	49.3	0.0	47.77	28.55	1.92	0.32

Panel B: Out-of-sample estimates

	FF5	FF6	FF5$_c$	FF6$_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	3.0	6.5	1.9	1.3	0.3	2.1	0.01	0.04	1.12	0.25
FF6	97.0	0.0	38.8	8.8	18.7	3.8	6.2	0.40	0.80	1.30	0.24
FF5$_c$	93.5	61.2	0.0	5.4	24.6	1.2	5.5	0.02	0.06	1.35	0.25
FF6$_c$	98.1	91.2	94.6	0.0	51.1	11.0	12.9	2.77	4.62	1.49	0.25
Q4	98.7	81.3	75.4	48.9	0.0	5.1	12.3	0.51	0.91	1.49	0.33
BS6	99.7	96.2	98.8	89.0	94.9	0.0	40.6	38.44	29.68	1.80	0.45
DHS	97.9	93.8	94.5	87.1	87.7	59.4	0.0	57.85	29.26	1.84	0.33

Table 10: **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 Figure 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 with: $(Variable - Mean)/(Max - Min)$. Illiquidity is calculated following Amihud (2002) and transaction costs following Hasbrouck (2009). Monthly characteristics are constructed using data from January 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). Asterisks are used to indicate significance at a 10% (*), 5% (**), or 1% (***) level.

	30-70	NYSE	BE	Micro	PRC	Utilities	Financials	Industry	VW	Independent	Jun_Size
Factor-	-1.13***	-0.16	-1.15***	-0.63***	-1.47***	-0.07	-0.17**	-0.29***	-0.39***	0.01	0.04
Exposure	(-4.33)	(-0.99)	(-4.67)	(-4.58)	(-4.78)	(-1.76)	(-2.71)	(-3.81)	(-4.21)	(1.02)	(1.36)
Illiquidity	-0.13**	-0.89***	-0.04**	1.99***	-1.95***	-0.08***	0.04*	-0.01	-1.69***	-0.04	-0.00
	(-2.46)	(-8.91)	(-2.46)	(5.44)	(-5.43)	(-7.97)	(2.06)	(-0.53)	(-5.74)	(-0.94)	(-0.03)
Transaction-	-0.06***	-0.15***	-0.00	0.11***	-0.10***	-0.01***	-0.01***	0.05***	-0.15***	-0.00***	-0.02***
Cost	(-5.36)	(-6.85)	(-1.08)	(7.48)	(-5.63)	(-9.29)	(-11.89)	(6.44)	(-6.85)	(-4.05)	(-5.83)

A Sorting Variables

This Appendix explains the sorting variables in more detail.

Market: Market is the return on the CRSP value weighted stock market index in excess of the risk-free rate.

Market capitalization: market capitalization is 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). The monthly updated book-to-market ratio is used to construct the monthly value factor (HML(m)).

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. The subtle difference is that for CMA, we filter observations with negative annual book equity, whereas for the IA factor, it is the quarterly book equity.

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

Quarterly Return-on-equity: This is the 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):

$$CAR_i = \sum_{d=-2}^{d+1} (R_{i,d} - Rm, d) \tag{7}$$

where $R_{i,d}$ 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 the earnings announcement date 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 also 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.

B The 2x3x3 Sorting Procedure

The results in the main body of the paper focus on 2x3 sorts, as in Fama and French (1993). In this appendix we consider 2x3x3 sorts. For example, the IA and ROE factors from Hou et al. (2015) are constructed using the 2x3x3 independent sorting procedure, as they independently sort on market capitalization, the annual change in total assets, and the quarterly return on equity.

It should be noted that there is not always clear guidance on which additional sort one should pick to include in 2x3x3 sorts. In the 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 a rationale to use the 2x3x3 sorting methodology. However, there is no theoretical guidance on how to construct FF-factor or DHS-factors using a 2x3x3 sort, or guidance which additional characteristic should be added in the 2x3x3 sort. This additional dimension leads to another degree of freedom, where the researcher has a wide range of options to select from.

Also note that using the 2x3x3 sorting methodology may lead to sparse or even empty portfolios. We construct an additional 2048 versions of each factor, using 2x3x3 sorting instead of 2x3 sorting, with the HML factor as the second sorting characteristic when constructing the FF and DHS factors. When an underlying portfolio of one of the leg is empty (say big-high-high), we consider the whole factor leg as missing. In Table D.1, we count for how many sets of construction choices (out of 2048) we obtain empty portfolios, in at least one month. When we construct 2x3x3 factors using 30-70 breakpoints, we find that most factors have no empty portfolio. The only exception is RMW, CMA, and PEAD, with 36, 12, and 32 sets of choices (out of 2048) missing data. Using 80-20 breakpoints limits the cross-section, and allows the occurrence of empty portfolios to increase. For RMW, we find that emptiness occurs in 400 sets of choices. As we mentioned before, a shortcoming of the independent sorting is that it may cause sparse portfolios. Panels C and D consider independent and dependent sorting, respectively. With 2x3x3 sorting, we indeed find that sparsity comes from the independent sorts. For dependent sorts, we never find empty portfolios. In Panel E and F, we show the sparsity for NYSE and NAN portfolios. Using NAN breakpoints creates a wider universe to select stocks from, making empty portfolios less likely when compared to using NYSE breakpoints. For NYSE we indeed find empty portfolios, ranging from 136 to 292 construction sets, whereas we do not find empty portfolios when NAN breakpoints are used.

Next, we consider the impact of 2x3x3 sorting jointly, with all other construction choices, on the Sharpe ratios of the Q factors (separately) and all factors (together). These results are shown in Figure D.1 and D.2, respectively. We find that using the 2x3x3 sorts increases the Sharpe ratio across all construction choices. For example, using NAN breakpoints and 2x3 sorts yields an average Sharpe ratio of 0.82, whereas 2x3x3 sorting yields a Sharpe ratio of 0.98 for Q-factors. Hence, the 2x3x3 sorting methodology is a construction choice that is able to consistently increase the risk-adjusted return of factors.

Finally, Table D.2 shows the amount of firms that are included in the Q-factor 2x3x3 portfolios. The 2x3x3 sorting methodology may stratify stocks into smaller segments where more extreme positions are overweighted. For example, the Big-High-High portfolio receives a weight of 1/6 in the High portfolios of the second and third characteristic. We find that the extreme portfolios typically contain less than 100 stocks. For example, Big-Low-Low contains an average of 51 stocks when excluding microcaps, and 34 when we use NYSE breakpoints.

C Net returns

Gross returns do not represent what is actually achievable by investors. To evaluate net returns, we estimate individual stock-level transaction costs as in Detzel et al. (2021) by using the estimation procedure from Hasbrouck (2009). Appendix C.1 explains how we estimate turnover and Appendix C.2 explains the estimation of transaction costs. Appendix C.3 examines net Sharpe ratios. Appendix C.4 focuses on model comparison with net returns.

C.1 Turnover

We estimate portfolio turnover for each factor. The turnover of an individual stock at time t ($TO_{i,t}$) is calculated by taking the absolute value of the difference between the portfolio weight at the start of the month ($W_{i,t}$) and the weight at the end of the past month ($W_{i,t-1,end}$). The turnover of the long leg of a factor is then defined as:

$$TO_{long,i,t} = \sum_{i=1}^{N_t} |W_{i,t} - W_{i,t-1,end}| \tag{8}$$

The turnover of the short-leg is defined in a similar way. The turnover of the long-short factor is defined as the sum of both the long and the short portfolios.

C.2 Transaction costs

We estimate transaction costs at the individual stock-level using the procedure of Hasbrouck (2009). This procedure yields effective spreads that highly correlate ($\geq$95%) with those from the high-frequency Trade and Quote (TAQ) database and allows for an estimation of effective spreads for public companies in the CRSP database using their daily price series. The procedure entails estimating transaction costs using a Bayesian-Gibbs sampler on the generalized stock price models of Roll (1984):

$$V_t = V_{t-1} + \epsilon_t \tag{9}$$

$$P_t = V_t + cQ_t \tag{10}$$

where V_t denotes the log midpoint of the prior bid-ask price (the "efficient price"), P_t denotes the log trade price (the "real price"), and Q_t indicates the sign of the last trade of the day. Q_t equals +1 for a buy, and -1 for a sale. ϵ_t is a random public shock to the efficient price V_t, and c is the effective one-way transaction cost. The above equations imply that:

$$\Delta P_t = \Delta c Q_t + \epsilon_t \tag{11}$$

Hasbrouck (2009) estimates c using an augmented version of the equation:

$$\Delta P_t = \Delta c Q_t + \beta R_{m,t} + \epsilon_t \tag{12}$$

where $R_{m,t}$ denotes the market return. Because the procedure from Hasbrouck (2009) yields missing observations, we impute observations by following the matching procedure from Detzel et al. (2021). First, for each stocks i on month t we compute:

$$M_i = \sqrt{(rank(ME_i) - rank(ME_j))^2 + (rank(IVOL_i) - rank(IVOL_j))^2} \tag{13}$$

where ME_i is the market capitalization and $IVOL_i$ is the 1-year idiosyncratic volatility estimate for firm i. If the transaction cost is missing for stock i on month t, we impute the transaction cost by finding the stock j that has the smallest difference between M_i and M_j, and using the transaction cost estimate of stock j.

For the long-leg, we compute portfolio-level effective spreads as follows:

$$TC_{long,t} = \sum_{i=1}^{N_t} |W_{i,t} - W_{i,t-1,end}| * c_{i,t} \tag{14}$$

where $c_{i,t}$ denotes the estimated transaction cost for stock i in period t. Transaction costs for the short-leg is defined similarly. Portfolio transaction costs for the long-short portfolios are equal to the sum of the transaction costs of each leg.

C.3 Net Sharpe Ratios

Figure D.3 shows the net Sharpe ratio distribution across sets of construction choices for each factor, based on value-weighted (Panel A) and equal-weighted (Panel B) net factor returns. In line with our findings on a gross basis, we also observe large variation in Sharpe ratios on a net basis. Some construction methods, for a given factor, yield negative net Sharpe ratios. The PEAD factor, for example, yields Sharpe ratios between -0.43 and 0.34 using value-weighting. The financing factor yields the highest average net Sharpe ratios, ranging between 0.28 and 0.74 when value-weighting.

Figure D.4 shows annualized maximum net Sharpe ratios by construction choice, averaged over factor models. Using 20-80 breakpoints yields a lower maximum Sharpe ratio (0.14) than 30-70 breakpoints (0.16), which could be explained by 20-80 breakpoints tilting towards extreme (small) stocks, which have higher transaction costs. Likewise, including microcaps yields lower net Sharpe ratios (0.12) than excluding microcaps (0.18). Including a price filter also improves the net Sharpe ratio (0.18 vs 0.12), since this excludes small illiquid stocks with high transaction costs. Furthermore, using value-weighting instead of equal-weighting puts less weight towards microcaps and yields an average Sharpe ratio of 0.20 versus 0.10. With gross returns, we documented that Sharpe ratios are higher when we include microcaps and use equal-weighting. With net returns, we thus find the opposite effect, due to the differential costs involved. Overall, our findings imply that construction choices also materially affect factor performance on a net basis.

C.4 Model comparison

Detzel et al. (2021) 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. Hence, we also consider net factor returns when reporting maximum Sharpe ratios on an annualized basis. For the mean-variance analysis with transaction costs, we follow the approach in Novy-Marx and Velikov (2016). More specifically, we estimate mean-variance optimal weights by using a long and short version of all the assets in the portfolio, net of transaction costs, subject to a no-shorting constraint on portfolio weights.

Figure D.5 shows the model selection results when we use net factor returns. Maximum Sharpe ratios decline using net returns, compared to the earlier presented gross returns. The value-weighted net FF6 model, with cash profitability, yields a net Sharpe ratio of 0.88. The BS6 model earns a net Sharpe ratio of 0.80. Using net returns, the DHS3 model yields the highest maximum Sharpe ratio (0.93), on average. We find that the average net maximum Sharpe ratio for the BS6 model is smaller than that of the DHS3 model when we use net returns instead of gross returns. We find that the BS6 model has the highest maximum Sharpe ratio in 1.7% of all construction sets, whereas this equals 85.7% for the DHS3 model.

Table D.3 reports the portfolio weights that correspond to the ex-post mean-variance efficient portfolios constructed from the candidate factor models using net returns, 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 weights are derived by adding a no-shorting constraint in the mean-variance analysis, following Novy-Marx and Velikov (2016). Across all models, we find that the average market weight increases relative to the results based on gross returns. Since transaction costs are incurred, factors are less profitable and more weight is allocated towards the market. Most factor weights decrease due to transaction costs. For example, CMA in the FF5 model decreases from 47.6% (gross) to 28.3% (net). In addition, due to the no-shorting constraint, low weights are allocated to factors with high transaction costs and negative net alphas. One example of such a case is the PEAD factor. It has a net weight of 6.8%, compared to a 58.1% gross weight, and a 9.6% standard deviation. For multiple construction choices, PEAD has a negative net alpha, thereby binding the no-short constraint and consequently receiving zero weight. The net DHS model predominantly consists of the financing factor (55.1%) and the market factor (38.1%).

Regarding the efficient frontier, Table D.4 shows the results when we focus on factor returns net of transaction costs. Due to these transaction costs, estimated alphas are closer to zero. Adding BS6 factors to FF models expands the efficient frontier between 0.03% and 0.08% per month, with standard deviations between 0.02% and 0.04%. Hence, there are construction methods for which the added value of the BS6 factors to the FF models is zero. The DHS factors improve FF models between 0.15% and 0.21% per month with standard deviations between 0.06% and 0.07%. Therefore, there are fewer construction methods that reach alphas closer to zero when adding the DHS model compared to the BS6 model. Again, our results imply that construction methods can influence model selection exercises, as shown by the relatively large standard deviations.

Table D.5 presents the results on economic significance using net returns. This exercise provides a more realistic view of the extent to which the investment opportunity set improves. We find that adding the Q4 factors improves the Fama-French models between 4.2% ($FF6_c$) and 15.4% with standard deviations between 4.7% and 10.9%, on average. For multiple construction methods, the Q4 factor adds little to no improvement relative to the FF models. Similarly, the BS6 and DHS factors improve FF factor models less compared to the analysis using gross returns, which is due to these models containing factors with relatively high turnover and transaction costs. Still, adding the BS6 factors to FF5 improves the Sharpe ratio by 42.8%, on average, with a standard deviation of 15.1%. Overall, we find that net economic significance varies due to differences in construction choices.

As a final analysis, we consider net returns for in-sample (Panel A) and out-of-sample (Panel B) estimation in Table D.6. It can be seen from Panel A that the 6-factor model with cash profitability is the best model in 20.6% of the cases, compared to 4.6% when using gross settings. Taking transaction costs into account, the BS6 model is no longer the model with the highest win-probability (15.4%). The DHS model has a win-probability of 57.1%. For out-of-sample estimates using net returns, the DHS model is also the model with the largest win-probability (71.2%).

D Additional Figures and Tables

Figure D.1: **Construction choices and gross Sharpe ratios for the Q-Factor Model.** This figure shows the impact of construction choices on the Sharpe ratio averaged over factors. Sharpe ratios are computed on a gross-basis and are annualized. The construction choice definitions are the same as in Figure 3. “233” (“23”) denotes that the factors are constructed using a 2x3x3 (2x3) sorting methodology. Monthly Q-factor returns are constructed using data from January 1972 to December 2021.

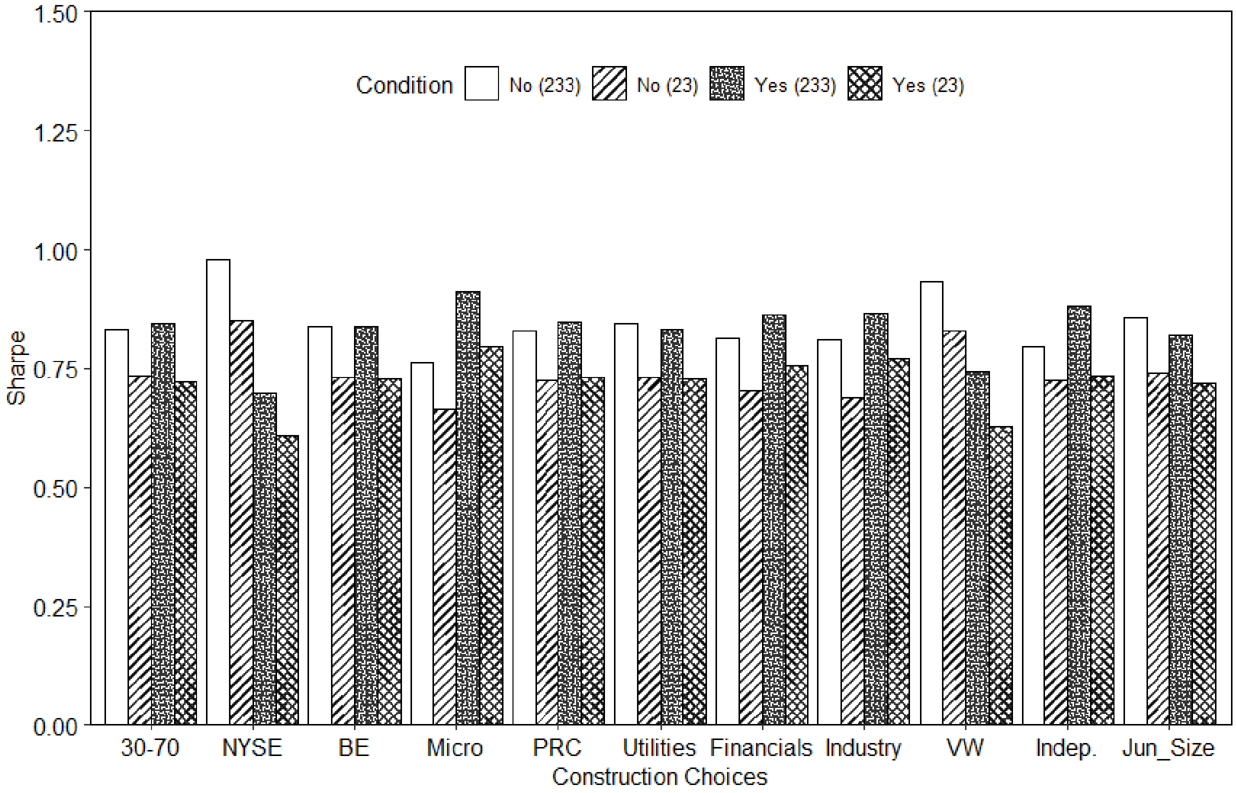


Figure D.2: **Construction choices and gross Sharpe ratios averaged over all factors.** This figure shows the impact of construction choices on the Sharpe ratio averaged over factors. Sharpe ratios are computed on a gross-basis and are annualized. The construction choice definitions are the same as in Figure 3. "233" ("23") denotes that the factors are constructed using a 2x3x3 (2x3) sorting methodology. Monthly Q-factor returns are constructed using data from January 1972 to December 2021.

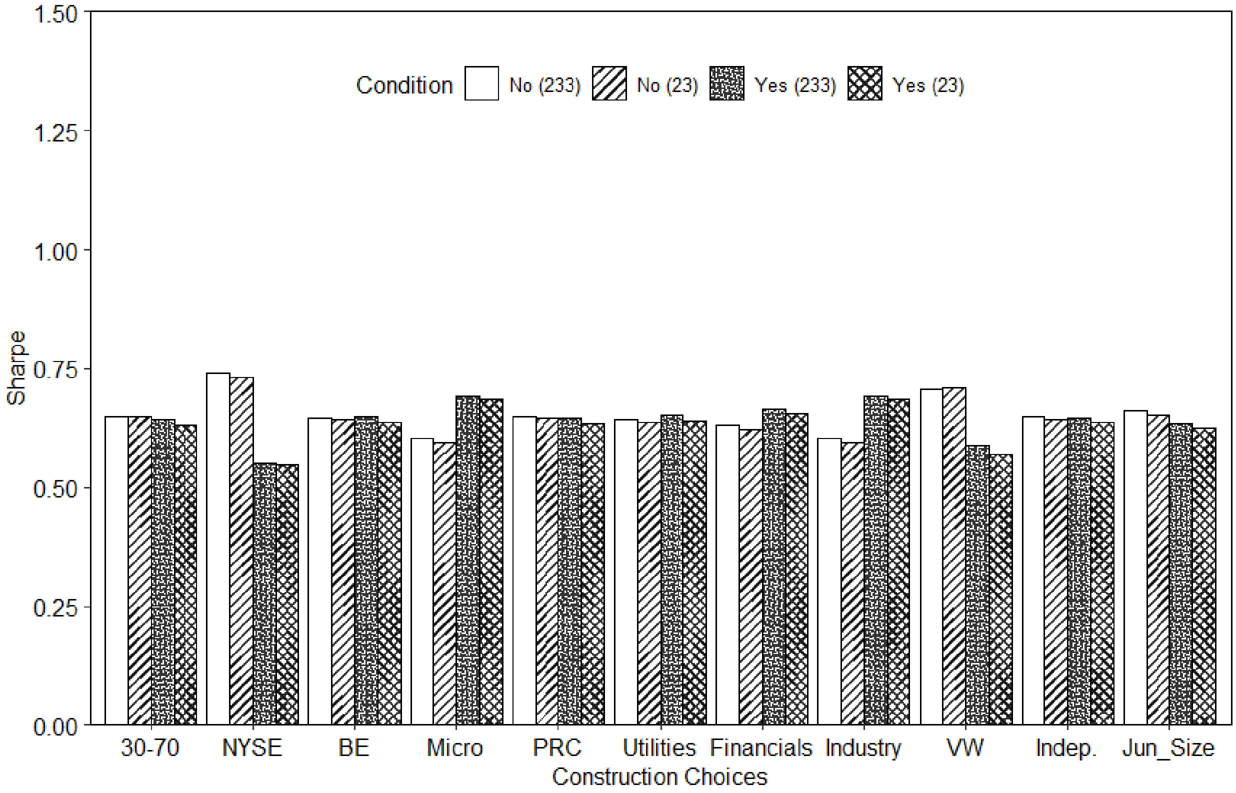
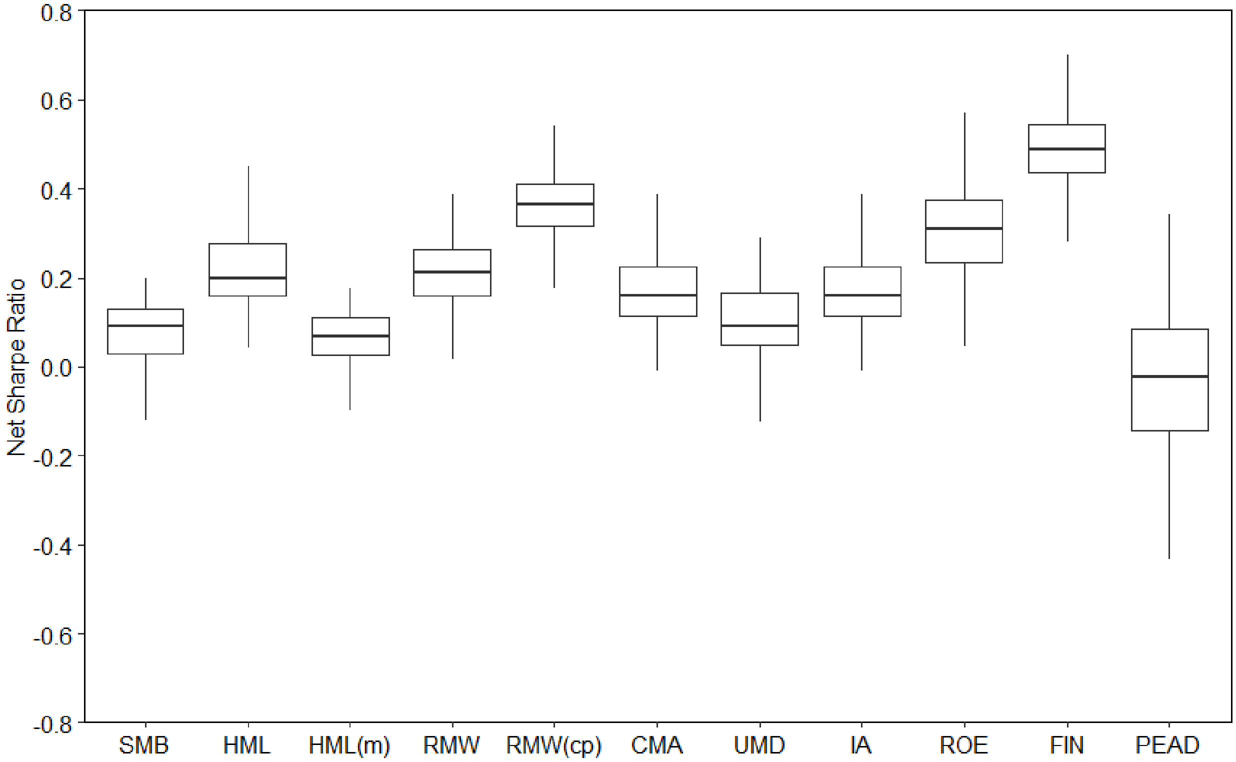
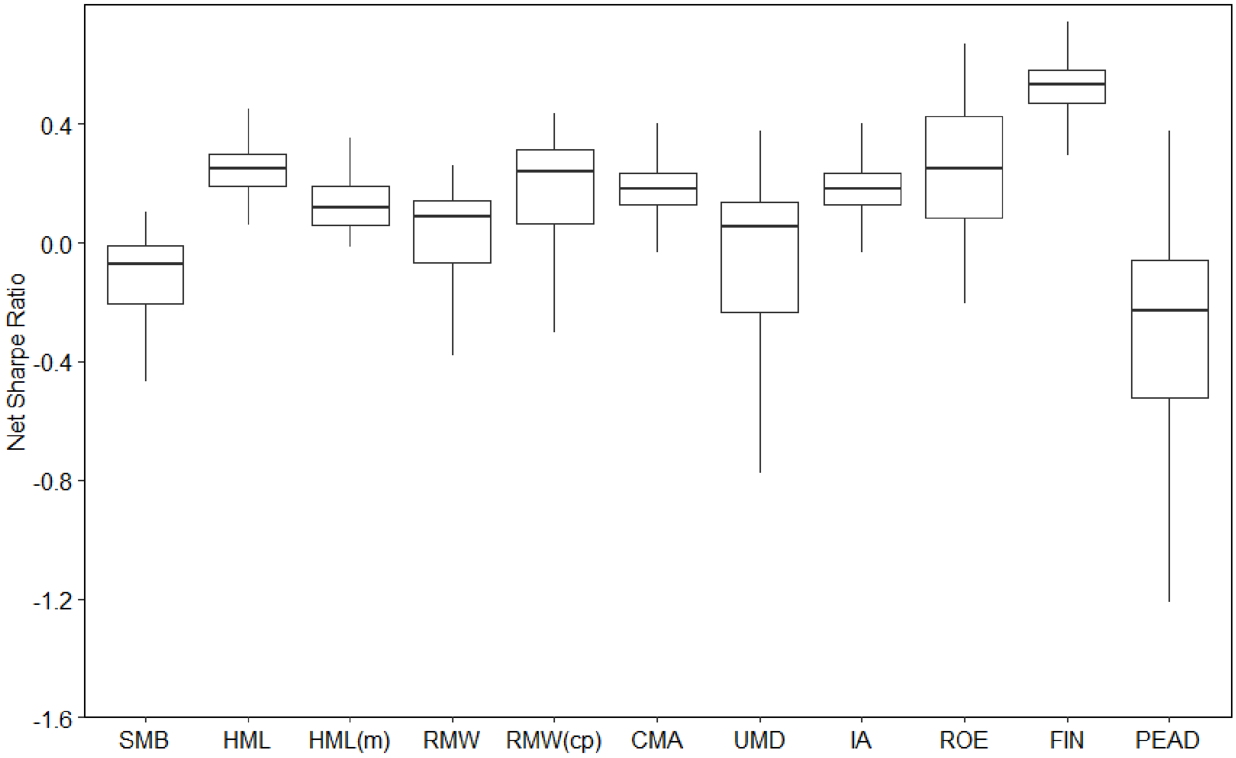


Figure D.3: **Sharpe ratio variation within factors net of transaction costs:** This figure plots the distribution of annualized value-weighted (subfigure A) and equal-weighted (subfigure B) Sharpe ratios for long-short factor returns net of transaction costs, where a factor is constructed N times by using the N 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, to calculate these Sharpe ratios, runs from January 1972 until December 2021.



(A) Value-Weighted

Figure D.3: **Sharpe ratio variation within factors net of transaction costs** – continued.



(B) Equal-Weighted

Figure D.4: **Construction choices and net Sharpe ratios.** This figure shows the impact of construction choices on the Sharpe ratio averaged over factors. Net Sharpe ratios are annualized. The construction choice definitions are the same as in Figure 3. Monthly factor returns are constructed using data from January 1972 to December 2021.

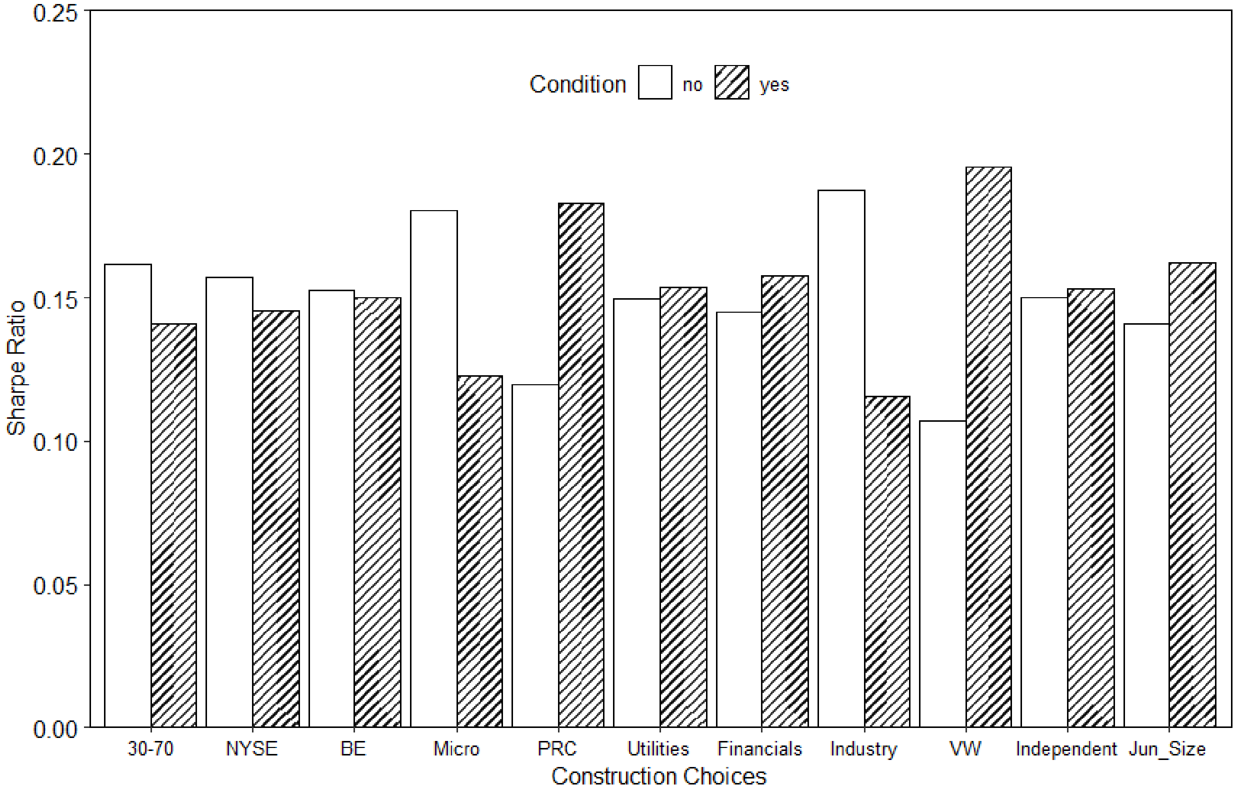


Figure D.5: **Selecting factor models using net returns.** This figure shows the maximum gross Sharpe ratio (annualized) from the factors from the factor models listed on the horizontal axis. The white bar shows the maximum Sharpe ratio obtained by using equal weighted factor returns. The dashed bar shows the maximum Sharpe ratio using value weighted factor returns. The error plot shows the variation in the maximum Sharpe ratios for a given factor model, across construction choices. The data runs from January 1972 until December 2021. The figure shows the results using net returns, taking transaction costs into account.

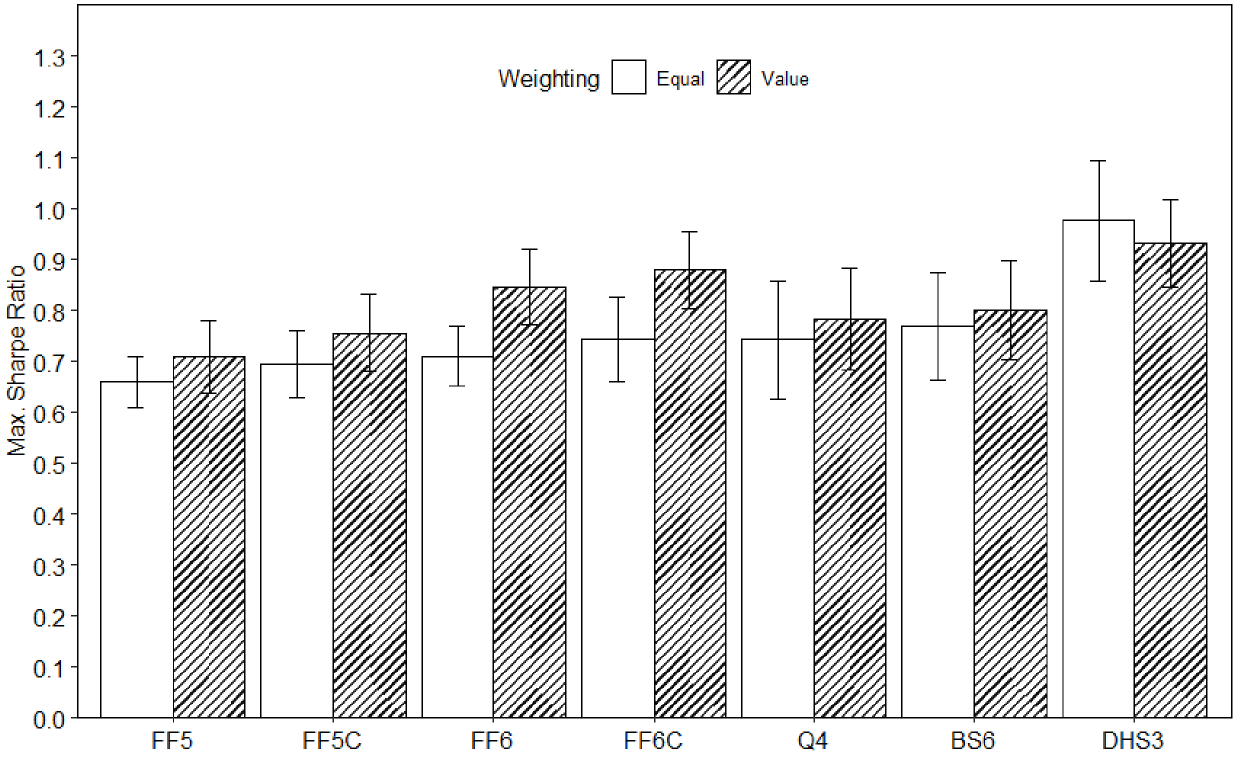


Table D.1: **Portfolio sparsity.** This table shows how many construction methods, for a given filter, contains at least one missing month of portfolio returns using a 2x3x3 sorting method. The first sorting characteristic is market capitalization (size), and the second sorting characteristic is the book-to-market ratio (value). The third sorting characteristic is listed in the first column.

	Panel A: 30-70				**Panel B: 80-20**				**Panel C: Independent**				**Panel D: Dependent**			
	2nd sort		**3rd sort**		**2nd sort**		**3rd sort**		**2nd sort**		**3rd sort**		**2nd sort**		**3rd sort**	
Factors	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
RMW	36	0	36	0	400	4	392	20	436	4	428	20	0	0	0	0
RMW$_{cp}$	0	0	0	0	256	0	240	16	256	0	240	16	0	0	0	0
CMA	12	0	12	0	124	0	124	0	136	0	136	0	0	0	0	0
MOM	0	0	0	0	136	16	28	124	136	16	28	124	0	0	0	0
PEAD	32	0	32	0	384	48	384	108	416	48	416	108	0	0	0	0
FIN	0	0	0	0	252	0	236	16	252	0	236	16	0	0	0	0

	Panel E: NYSE				**Panel F: NAN**				**Panel G: Incl. Micro**				**Panel H: Ex. Micro**			
	2nd sort		**3rd sort**		**2nd sort**		**3rd sort**		**2nd sort**		**3rd sort**		**2nd sort**		**3rd sort**	
Factors	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
RMW	292	4	292	4	0	0	0	0	220	0	212	16	216	4	216	4
RMW$_{cp}$	208	0	208	0	0	0	0	0	148	0	132	16	108	0	108	0
CMA	136	0	136	0	0	0	0	0	32	0	32	0	104	0	104	0
MOM	128	16	28	116	0	0	0	0	72	16	8	68	64	0	20	56
PEAD	288	48	288	48	0	0	0	0	224	0	224	40	192	48	192	68
FIN	236	0	236	0	0	0	0	0	128	0	112	16	124	0	124	0

Table D.2: **Firms in Q factor portfolios.** This table shows the average number of positions for Q-factor portfolios, averaged across all 2048 construction methods. B and S denotes the Big and Small portfolio, respectively. H and L (second letter) denotes High and Low for the IA characteristic, whereas the third letter denotes High or Low for the ROE characteristic. The ROE and IA factor are constructed using data from January 1972 to December 2021.

Choice	BHH	BHN	BHL	SHH	SHN	SHL	BLH	BLN	BLL	SLH	SLN	SLL
Ex. Micro	72	103	50	68	118	78	52	92	51	51	105	88
Incl. Micro	90	126	59	104	185	142	58	103	57	82	167	180
Dep.	65	117	65	89	159	103	60	109	60	77	138	93
Ind.	97	112	44	83	145	117	50	85	48	56	134	176
20-80	51	114	35	56	152	77	34	94	33	43	137	96
30-70	111	115	74	116	151	143	76	100	75	90	135	172
NAN	105	149	70	71	128	86	75	135	73	66	140	121
NYSE	57	79	39	101	175	134	35	60	34	67	132	147

Table D.3: **Mean-variance efficient portfolio weights using net returns.** 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 net of transaction costs. The sample period is from January 1972 to December 2021.

	Mkt	SMB	HML	RMW	CMA	UMD	RMW$_{cp}$	IA	ROE	HML$_d$	FIN	PEAD
FF5	33.4	1.4	19.2	17.6	28.3							
	(6.0)	(2.6)	(17.2)	(11.9)	(12.5)							
FF6	30.3	1.5	21.6	15.9	20.8	10.0						
	(6.1)	(2.5)	(15.1)	(10.5)	(11.7)	(5.9)						
FF5C	28.7	3.7	13.3		35.4		19.0					
	(6.5)	(4.3)	(13.5)		(16.1)		(10.3)					
FF6C	27.0	3.5	15.4		31.9	14.9	7.3					
	(6.7)	(4.1)	(12.6)		(14.5)	(9.8)	(4.9)					
Q4	28.3	3.4						40.2	28.2			
	(4.7)	(3.6)						(6.6)	(7.7)			
BS6	25.9	2.8					5.9	29.6	26.0	9.8		
	(3.7)	(3.2)					(5.6)	(9.3)	(6.8)	(6.8)		
DHS3	38.1										55.1	6.8
	(5.1)										(8.0)	(9.6)

Table D.4: **Frontier expansion using net returns.** 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 net returns, taking transaction costs into account. The data runs from January 1972 until December 2021.

Net	**Union Model (M1)**						
Base Model (M0)	FF5	FF5$_c$	FF6	FF6$_c$	Q4	BS6	DHS3
FF5	0.00	0.05	0.03	0.07	0.07	0.08	0.21
	(0.00)	(2.38)	(1.43)	(2.82)	(2.71)	(2.98)	(4.73)
	[0.00]	[0.03]	[0.03]	[0.04]	[0.04]	[0.04]	[0.06]
FF5$_C$	0.00	0.00	0.02	0.02	0.04	0.05	0.16
	(0.01)	(0.00)	(1.29)	(1.29)	(1.80)	(2.28)	(4.21)
	[0.00]	[0.00]	[0.02]	[0.02]	[0.03]	[0.03]	[0.07]
FF6	0.00	0.04	0.00	0.04	0.05	0.05	0.19
	(0.00)	(2.35)	(0.00)	(2.35)	(2.19)	(2.35)	(4.62)
	[0.00]	[0.02]	[0.00]	[0.02]	[0.04]	[0.03]	[0.06]
FF6$_c$	0.00	0.00	0.00	0.00	0.02	0.03	0.15
	(0.01)	(0.00)	(0.01)	(0.00)	(1.31)	(1.53)	(4.11)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.02]	[0.02]	[0.06]
Q4	0.02	0.04	0.03	0.05	0.00	0.02	0.18
	(1.18)	(2.03)	(1.49)	(2.33)	(0.00)	(1.07)	(4.38)
	[0.02]	[0.03]	[0.02]	[0.03]	[0.00]	[0.01]	[0.05]
BS6	0.01	0.04	0.01	0.04	0.00	0.00	0.16
	(1.01)	(2.01)	(1.01)	(2.01)	(0.00)	(0.00)	(4.31)
	[0.02]	[0.03]	[0.02]	[0.03]	[0.00]	[0.00]	[0.05]
DHS	0.02	0.03	0.03	0.04	0.04	0.04	0.00
	(0.88)	(1.54)	(1.48)	(2.00)	(1.82)	(2.06)	(0.00)
	[0.02]	[0.03]	[0.02]	[0.03]	[0.03]	[0.03]	[0.00]

Table D.5: **Economic significance using net returns.** 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 net returns, taking transaction costs into account. The standard deviation of the increase in Sharpe, across construction methods, is reported. The data runs from January 1972 until December 2021.

Net	**Union Model (M1)**						
Base Model (M0)	FF5	$FF5_c$	FF6	$FF6_c$	Q4	BS6	DHS3
FF5		13.8	5.9	18.7	15.4	17.9	42.8
		(9.5)	(4.6)	(11.1)	(10.9)	(10.3)	(15.1)
$FF5_c$	0.0		4.2	4.2	6.8	9.4	29.4
	(0.1)		(3.9)	(3.9)	(7.0)	(6.6)	(12.7)
FF6	0.0	11.9		11.9	10.2	11.3	37.3
	(0.0)	(8.1)		(8.1)	(8.3)	(7.8)	(13.0)
$FF6_c$	0.0	0.0	0.0		4.2	5.0	26.2
	(0.1)	(0.0)	(0.1)		(4.7)	(4.7)	(11.2)
Q4	3.6	9.2	4.9	11.2		3.1	32.1
	(3.7)	(7.5)	(3.7)	(8.2)		(3.1)	(11.1)
BS6	2.8	8.7	2.8	8.7	0.0		29.5
	(3.3)	(7.7)	(3.3)	(7.7)	(0.0)		(9.9)
DHS3	2.1	5.1	4.0	6.9	5.0	6.2	
	(2.9)	(5.4)	(3.3)	(5.8)	(4.3)	(4.4)	

Table D.6: **In-sample and out-of-sample Sharpe ratios using net returns.** 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. We use the factor models listed in Table 1. "SR" reports the maximum Sharpe ratio of the row model, averaged across construction methodologies. $\sigma(SR)$ reports the standard deviation of the maximum 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 estimates and Panel B shows the out-of-sample estimates using net returns. The estimates are based on 100.000 in-sample and out-of-sample simulation runs. Each simulation run splits the 600 sample months, running from January 1972 until December 2021, into 300 adjacent pair-months. The run randomly draws a sample of pairs (with replacement). The in-sample simulation randomly draws one month from each pair within a run. The remaining months form the out-of-sample. The in-sample observations are used to calculate in-sample Sharpe ratios and portfolio weights. The in-sample portfolio weights are applied to the out-of-sample returns to produce an out-of-sample Sharpe ratio estimate.

Panel A: In-sample estimates (net returns)

	FF5	FF6	FF5$_c$	FF6$_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	3.4	0.0	1.8	30.5	14.6	13.2	0.00	0.00	0.78	0.07
FF6	83.7	0.0	56.9	0.0	56.7	38.5	26.6	0.00	0.00	0.87	0.10
FF5$_c$	76.2	33.0	0.0	4.2	50.6	25.9	21.4	0.57	1.31	0.84	0.09
FF6$_c$	88.2	73.5	83.0	0.0	71.7	54.1	35.4	20.61	19.88	0.93	0.12
Q4	69.0	42.9	49.1	28.0	0.0	0.0	21.3	0.00	0.00	0.85	0.11
BS6	85.2	61.3	73.8	45.7	92.6	0.0	31.5	15.45	10.73	0.91	0.10
DHS	86.8	73.4	78.6	64.6	78.7	68.5	0.0	57.12	21.33	1.00	0.11

Panel B: Out-of-sample estimates (net returns)

	FF5	FF6	FF5$_c$	FF6$_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	10.3	23.4	13.4	27.9	26.5	8.0	0.29	0.43	0.62	0.07
FF6	76.8	0.0	65.5	24.6	52.8	50.4	18.7	4.66	5.32	0.72	0.10
FF5$_c$	52.8	24.4	0.0	10.3	34.4	29.3	10.0	0.36	0.60	0.65	0.08
FF6$_c$	76.5	49.0	77.0	0.0	57.8	56.0	21.0	8.92	10.47	0.74	0.11
Q4	71.5	46.9	65.2	41.9	0.0	41.9	15.6	5.25	5.05	0.71	0.12
BS6	73.4	49.5	70.4	43.8	50.8	0.0	16.1	4.86	4.40	0.71	0.11
DHS	92.0	81.3	90.0	79.0	84.4	83.9	0.0	71.21	19.04	0.92	0.11