Mind Your Sorts

Amar Soebhag[1] and Bart Van Vliet[1,2]

[1]Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam 3000 DR, Netherlands.

[2]Robeco Quantitative Investing, Weena 850, 3014 DA Rotterdam.

May, 2022

Abstract

Factor returns obtained from characteristic-based sorting are based on their signal, but also on the procedure and choices of the factor construction method. Since there is no consensus on which choices to make when constructing factors, researchers face a number of degrees of freedom, potentially allowing for p-hacking. We focus on a wide-range of construction choices that are widely used in financial research on cross-sectional equity factors, and on purpose data mine over 250 different versions of each factor in our sample. We find that Sharpe ratios exhibit large and significant variation within a factor due to construction variation. Consequently, this variation impacts model selection exercises. We attribute the variation in factor performance, across construction choices, to variation in liquidity, factor breadth, and diversification.

JEL Classification: G11, G12, G15

Keywords: p-hacking, data-mining, equity factors, asset pricing models, portfolio management, factor investing, portfolio construction

A. Soebhag: soebhag@ese.eur.nl (corresponding author) and B.P. van Vliet b.p.vanvliet@ese.eur.nl. Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3000 DR, Netherlands. Financial support by Tinbergen Institute and Erasmus Research Institute of Management is gratefully acknowledged. We thank Guido Baltussen and Patrick Verwijmeren for fruitful discussions and feedback. The views expressed in this paper are not necessarily shared by Robeco Institutional Asset Management. We welcome comments, including references to related papers we have inadvertently overlooked.

1 Introduction

Characteristic-based portfolio sorting is a widely used procedure in modern empirical finance. It has been deployed to test theories in asset pricing, construct a wide range of pricing anomalies, and identify investment strategies that are profitable. The empirical applications of this method is too vast to list, and hundreds of factors have been identified using portfolio sorting. By now, the academic literature in finance documents over 450 factors that "supposedly" price the cross-section of equity returns. This growing list of factors is often referred as the "factor zoo". Harvey, Liu, and Zhu (2016) argue that almost all of the previous studies fail to take the multiple testing problem into account: some will appear to be "significant" by luck. The authors also note that journals incentivize researchers to find strong positive results. Such incentive might induce researchers to engage in p-hacking.

Many forms of p-hacking exist, some more serious or more difficult to spot than others. One form of p-hacking is to cherry-pick results, i.e. only report the factors that meet statistical hurdles and appear to be the most significant. P-hacking might also occur when a researcher conducts several statistical methods, but only reports the significant results from a subset of methods. Another form is the exclusion, selection, and manipulation of data. Researchers may opt to exclude outliers, select specific sample periods or implement certain data transformations. If a researcher makes these choices with the intention of producing a result that has a higher t-statistic, this is also considered p-hacking. The bottom-line is that researchers exhibit a large number of "degrees of freedom" regarding research choices (sample selection, method selection and so forth).

In this paper, we shed light on a new dimension of p-hacking that should be seriously considered by researchers, namely the freedom of choice pertaining factor construction methodologies. This ambiguity regarding construction choices creates an open playing field for researchers to construct factors in such a way that it may maximize some statistical criteria, such as maximizing Sharpe ratios and t-statistics. Hence, construction choices matter and may materially impact results. The empirical literature shows no consensus regarding factor construction methodologies, as implied by the wide variety of construction methods deployed in these studies. Differences in performance between factors might not only stem from signals, but are also driven by the fact that different construction schemes are deployed. To illustrate, Fama and French (2015) uses 2x3 portfolio sorts to construct factors, whereas Hou, Xue, and Zhang (2015) use 2x3x3 portfolio sorts. Typically, in model selection studies, factor models are compared amongst each other while using the original factor construction methodology. Selecting models with factors using different construction methods might bias conclusions in favour of asset pricing models that use favourable construction choices. Thus, construction choices can also materially impact model selection exercises. A fair model comparison can only be conducted when factors have a uniform construction method.

Which construction choices can be made by a researcher? The list of choices is abundant. As our first contribution, in this paper we restrict ourselves to a wide-range of popular binary construction choices, and we construct factors using each possible combination of choices. Specifically, we focus on 8 choices, leading to 256 (2^8) construction combinations. We consider choosing between including and excluding financials (1), including and excluding microcaps (2), NYSE or "NAN" (NYSE,AMEX,NASDAQ) breakpoints (3), industry neutralization or not (4), value-weighting and equal-weighting (5), independent vs. dependent sorts (6), 70/30 or 80/20 breakpoints (7), and sorting on the most recent market capitalization or from June (8). Second, we investigate how construction choices affect the performance of individual factors. Third, we show to what extent construction choices can affect model selection exercises. Fourth, we discuss the potential impact of construction choices on portfolio efficiency, and consider other aspects such as factor exposure, liquidity, and portfolio diversification. We do have to stress out that the aim of our paper is not to advocate a specific construction methodology, but rather to raise awareness about the consequences of using only one specific construction methodology. Preferably, to prevent data-mining, multiple construction methods should be considered (analogous to using a multiple hypothesis testing setting). Our sample consists of U.S. common stocks from CRSP and Compustat with data spanning from January 1972 until December 2021. Using this data, we construct widely-known factors and factor models, with in total 256 different versions of portfolio construction methods. Following Detzel et al. (2021), we focus on both gross and net returns. Transaction costs are estimated using Bayesian-Gibbs sampling as explained in Hasbrouck (2009).

Figure 1 shows the key result of our paper: factors exhibit large variation in Sharpe ratios in our set of possible construction methods. Hence, factor returns are not only a function of their sorting characteristic, but also a function of their construction choices. For example, the gross Sharpe ratio (annualized) of the canonical value factor of Fama and French (1993) may vary between 0.18 and 0.88, depending on how it is created. This variation in Sharpe ratio within factors, also impacts model selection exercises, since maximum Sharpe ratios of factor models also vary across construction methods. On a gross basis, the factor model of Barillas et al. (2020) has the largest maximum Sharpe ratio, but also varies heavily across construction methods. On a net basis, however, the behavioural-factor model of Daniel, Hirshleifer, and Sun (2020) has the largest net maximum Sharpe ratio, and is the least volatile across construction methods. Furthermore, we find that mean-variance weights and economic significance (much how augmenting a model by another model) are sensitive to construction methods. Lastly, we attribute the differences in factor performance, across construction methods, to differences in factor breadth, liquidity and diversification. For example, equal weighting results into portfolios with higher illiquidity and more spread in factor characteristic, consequently resulting in higher risk-adjusted gross returns relative to value-weighting.

Our paper contributes to several strands of the empirical asset pricing literature. First, it is related to the voluminous literature regarding p-hacking. Recent studies regarding data mining and market anomalies include Harvey et al. (2016), McLean and Pontiff (2016), Yan and Zheng (2017), Linnainmaa and Roberts (2018), and Hou, Xue, and Zhang (2020). The latter study constructs a database of 452 anomalies using a common set of factor construction procedures. This common set consists of choices that can be considered as "good practice". Similar to their study, we also compare factor models on a "apples-to-apples" basis, however we take an agnostic view about construction choices and consider a wide range of construction sets. Hou et al. (2020) find that many anomalies are driven by small stocks, which only consists of 3% of the market capitalization. Once factor portfolios are value-weighted, around two-third of the anomalies disappear. Giglio, Liao, and Xiu (2021), and Harvey and Liu (2020) further develop methods to control for the multiple testing problem. The p-hacking literature recognizes the randomness in data samples by taking into account that some factors will appear significant by "luck". Another way to interpret our paper is to think in terms of non-standard errors as introduced by Menkveld et al. (2021): in addition to a data-generating process there also exists an evidence-generating process (EGP), which translates the sample into evidence. This EGP exposes variation across researchers, resulting into a second layer of uncertainty, namely the non-standard error. Under this notion, our paper can be thought as modelling N hypothetical researchers who each independently construct factor returns using random construction methods, and subsequently calculate the Sharpe ratio. The non-standard error in our case would be the standard deviation for the generated Sharpe ratios across researchers (i.e. construction methods). Feng, Giglio, and Xiu (2020), and Kozak, Nagel, and Santosh (2020) use regularization techniques to tame the "factor zoo". This strand of literature focuses on how to deal with the large number of firm characteristics that have been documented to explain the cross-section of expected returns. Our paper differs by rather asking: how to deal with the large number of possible ways in which a factor can be constructed?

Second, this paper adds to the literature on evaluating factor models. Barillas and Shanken (2017) show that for models with traded factors, the extent to which each model is able to price the factors in the other model is what matters for model comparison, not the test assets. They propose the use of the squared maximum Sharpe ratio as a model comparison metric. Fama and French (2018) use the maximum squared Sharpe ratio to evaluate their 3-factor, 5-factor, and 6-factor models. Barillas et al. (2020) compares a range of models using the maximum squared Sharpe ratio, and finds 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. Detzel et al. (2021) show that failing to account for transaction costs materially impacts inferences when evaluating asset pricing models, biasing tests in favor of those employing high cost factors. However, the factors in the latter two studies are compared while different construction methods are used (2x3x3 Q factors versus 2x3 Fama-French factors). The aim of this paper is not to advocate one specific factor construction method, but rather to illustrate how factor construction choices impact factor returns and simple model selection exercises. We compare models by explicitly constructing factors using the same set of construction choices.

Third, some factor construction choices have been questioned by the academic literature. However, the focus is typically limited to a few choices, and are especially studied in isolation. Hou et al. (2019) show that the performance of factors is sensitive to breakpoints that are used to construct factors. Hou et al. (2020), in a large replication study, shows that many anomalies disappear when microcaps are excluded. A drawback of using independent sorts is that it potentially can produce empty portfolios. Hence, in studies where data is scarce or in an international empirical asset pricing context, the use of dependent sorts have been advocated instead of independent sorts. Some examples of studies that use dependent sorts are Ang et al. (2006), Novy-Marx (2013), Wahal and Yavuz (2013). Daniel et al. (2020) question the standard construction method that asset pricing scholars employ to form characteristics-based factors. The authors note that the characteristic, used to form a factor, picks up the variation in the loading with respect to both the priced risk factor as well as unpriced sources. Hence, the factor is no longer mean-variance efficient. Consequently, they propose to hedge out such unpriced risks, and propose industry as one component that should be hedged out. Our study comprehensively studies eight construction choices, both in isolation as well as jointly, and is the first in the literature to our knowledge.

The remainder of this paper is organized as follows. We describe the factor data and portfolio construction process in section 2. Section 3 shows how different factor construction methods may materially impact model selection exercises. Section 4 shows how different construction methods affects several key portfolio characteristics. In section 5, we zoom in on the 2x3x3 versus the 2x3 sorting procedure. Section 6 concludes.

2 Constructing Factor Models

2.1 Data

Monthly return and prices for U.S. equities are obtained from the Center for Research in Security Prices (CRSP), and accounting information is retrieved from the Compustat Annual and Quarterly Fundamental Files. Our sample consists of stocks listed on the NYSE, AMEX, and Nasdaq stock with share codes 10 or 11 (i.e. all common stocks). The sample period spans January 1972 to December 2021, thereby covering 600 months of factor returns. The starting date is restricted by the availability of the quarterly earnings announcements dates (to construct the PEAD factor), and quarterly book equity data (to construct the ROE factor). Firms with negative book equity are excluded. Definitions of the sorting variables are provided in the Appendix.

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. 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, whilst at the same time taking a short position in the Small.Low and Big.Low portfolio[1].

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

[1] An exception for this methodology is the Small minus Big factor (SMB). The SMB portfolio is created with a multi-factor perspective, it is long the small portfolios of all of the factors and short the big portfolios of all the factors.

Furthermore, we estimate portfolio turnover for each factor. The turnover of an individual stock at time t ($TO_{long,i,t}$) 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 ending weight of the past month ($W_{i,t-1,end}$):

$$TO_{long,i,t} = \sum_{i=1}^{N_t} |W_{i,t} - W_{i,t-1,end}| \tag{2}$$

Note that if a stock changes from the Big.High to the Small.High portfolio, and $W_{i,t}$ in the Small.High portfolio is equal to the $W_{i,t-1,end}$ in the Big.High portfolio, the stock had an effective turnover of 0. In addition, at the start of the period every portfolio has an effective turnover of 100% as the summed portfolio weights ($\sum_{i=1}^{N_t}, W_{i,t-1,end}$) are 0. We calculate the turnover of the long-short factor as the sum of both the long and the short portfolios.

Lastly, we estimate the transaction costs at the individual stock-level by following Detzel et al. (2021), and using the estimation procedure from Hasbrouck (2009). This procedure yields effective spreads that highly correlate (95%+) with those from the high-frequency Trade and Quote (TAQ) database. In addition, this procedure allows to estimate effective spreads for public companies in the CRSP database, using their daily price series. Specifically, we estimate transaction costs using a Bayesian-Gibbs sampler on the generalized stock price models of Roll (1984):

$$V_t = V_{t-1} + \epsilon_t \tag{3}$$

$$P_t = V_t + cQ_t \tag{4}$$

Where V_t denotes the log midpoint of the prior bid-ask price, which is also called the "efficient price". P_t denotes the log trade price, called the "real price". 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. Equation (3) and (4) imply that:

$$\Delta P_t = \Delta c Q_t + \epsilon_t \tag{5}$$

Hasbrouck (2009) estimates c using an augmented version of equation (5):

$$\Delta P_t = \Delta c Q_t + +\beta R_{m,t} + \epsilon_t \tag{6}$$

Where $R_{m,t}$ denotes the market return. The procedure from Hasbrouck (2009), however, yields missing observations. We impute these observations by following the matching procedure from Detzel et al. (2021). Specifically, in month t, we assign to each stock i with missing c's that of stock j with the closest distance in the market cap - idiosyncratic space. We assign to stock i the estimated spread of the stock j with the smallest value of:

$$\sqrt{(rank(ME_i) - rank(ME_j))^2 + (rank(IVOL_i) - rank(IVOL_j))^2} \tag{7}$$

We use the estimated effective spreads of individual stocks to proxy the transaction costs made within factor portfolios, as follows:

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

Where $c_{i,t}$ denotes the estimated transaction cost for stock i in period t.

Table 1 reports the summary statistics of the factors that we include in our sample. The reported statistics are averaged over the set of construction methods. Gross value-weighted factor returns range between 0.17% (SMB) and 0.71% (UMD) per month, with monthly Sharpe ratios ranging between 0.18 (SMB) and 1.19 (PEAD). Gross returns and Sharpe ratios for equal weighted factors tend to be higher, due to overweighting towards small caps. Taking transaction costs into account leads to a decline in the Sharpe ratio for all factors. Value-weighted factors have net Sharpe ratios ranging from -0.07 (PEAD) and 0.49 (FIN). Especially high turnover factors, which rebalance monthly, such as momentum and PEAD, are expensive to trade leading to a low (or even negative) net return. Transaction costs are also higher for equal-weighted factors compared to value-weighted factor, hence equal-weighted factor portfolios are more expensive to trade.

2.2 Candidate factor models

We use multiple candidate factor models coming from Fama and French (2018) (and its variants), Hou et al. (2015), Daniel et al. (2020), and Barillas and Shanken (2018). The five-factor model of Fama and French (2015) uses the market, size, value, profitability and investment factors (which we abbreviate as FF5). FF5 can also be augmented by momentum, with the resulting model being abbreviated as FF6. Fama and French (2018) suggest to replace the profitability factor by a cash-profitability factor, with the corresponding FF5 and FF6 models being abbreviated as $FF5_c$ and $FF6_c$. Hou et al. (2015) consists of the market factor, size factor, a growth in book assets factor, and return on equity factor. We abbreviate this model as the "Q" factor model. Daniel et al. (2020) constructs a three-factor model consistent of the market factor, financing factor, and the post-earnings announcement factor, abbreviated as "DHS". Lastly, Barillas and Shanken (2018) combines 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.

Table 2 summarizes the factors underlying the factor models, and their key construction choices as implemented in their original studies. The Fama-French 5-factor model consist of the market factor, the size (SMB) factor, the value (HML) factor, the operating profitability (RMW) factor, and the investment (CMA) factor. 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 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. The UMD factor differs only in the rebalancing, which is monthly. In addition, we construct a cash-based version of both models, which only replaces operating profitability by a cash profitability factor. The Q4 factor model of (Hou et al., 2015) consists of the market factor, size (SMB) factor, investment (IA) factor, and return on equity (ROE factor). In the original set-up, these factors are derived from a 2x3x3 independent sort. The IA and ROE factor are rebalanced monthly. The BS6 model of Barillas et al. (2020) consists of the market factor, size (SMB) factor, a monthly updated value factor (HML(m)), the momentum (UMD) factor, and the investment (IA), and return on equity (ROE) factor). Lastly, we construct the financing (FIN) and post-earnings announcement drift (PEAD) factor from Daniel et al. (2020). Both the FIN and PEAD factor use 20-80 breakpoints in the characteristic dimension. The PEAD factor is rebalanced monthly.

2.3 Construction choices

We assert that factor returns depend on two things: its signal s_t (1), and a set of construction choices (Ω_i). Hence, when a factor is constructed, the return is always conditional on the construction choices: $r_t(s_t|\Omega_i)$. Each choice ω_j is a part of set of choices Ω_i, and different choices results in different sets of construction choices. The aim of this paper is to study the impact of Ω_i on factor returns. The list of construction choices ω_j is numerous, hence we focus on eight construction choices that are often made in the empirical finance literature.

These eight choices result in 256 (2^8) sets of construction choices, hence for each factor and factor model we also have 256 different versions. In this section, we explain which choices we include in our study, and link this to previous literature.

2.3.1 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 à la Fama and French (1993). First, stocks are sorted by their market capitalization, whereby stocks are split into "small" and "big" classifications based on the NYSE median break-point. Second, and independently, stocks are sorted on their characteristic, whereby stocks are classified into "high" and "low" based on the 30th and 70th percentile (calculated over the NYSE universe) of the characteristic. The intersection of these classifications result into six portfolios, from which the high-minus-low portfolio is derived.

The 30th and 70th percentile is used often in the characteristic dimension, for example in Fama-French models (Fama & French, 2018) and in the Q factor model (Hou et al., 2015). On the other hand, others have chosen to deploy the 20th and 80th percentile to sort portfolios in the characteristic dimension. Recent examples are McLean and Pontiff (2016), Stambaugh and Yuan (2017) and Daniel et al. (2020). The consequence of using the latter choice is that stocks with more extreme characteristics are selected into portfolios, increasing the spread in factor characteristic. If expected returns linearly increase in those stock characteristics, such factor portfolios will tend to have higher average returns. As such, factors that are constructed from different break-point rules should not be compared directly to each. We construct different versions of factors, where we either use the 30th-70th or 20th-80th breakpoint in the characteristic dimension.

2.3.2 Breakpoints Universe:

Another common practice is to calculate breakpoints over the NYSE universe. However, many studies also calculate breakpoints over the NYSE-AMEX-Nasdaq ("NAN") universe. Stambaugh and Yuan (2017) not only opt for the 20th and 80th percentile, they also use those over the NAN universe when forming portfolios. McLean and Pontiff (2016) find that the average return spreads of 97 anomalies decline out of sample and post publication, but the tests are based on NAN breakpoints. Yan and Zheng (2017) construct 18.000 fundamental signals, and construct high-minus-low deciles using NAN breakpoints, as opposed to NYSE breakpoints. Using NAN breakpoints, compared to NYSE breakpoints, tilts portfolios more towards small stocks. We show this in figure 3 for our CRSP sample, where we plot the median size breakpoints using the NYSE and NAN universe. The log of the market capitalization is always higher for the NYSE universe compared to the NAN universe. Furthermore, using NAN breakpoints also tilts towards more extreme factor characteristics. Having smaller and more extreme positions in portfolios may exaggerate anomaly profits. Since both choices often occur in the literature, we also include this choice in our set of choices.

Figure 3 shows the effect using NYSE or NAN breakpoints on the size dimension of the CRSP universe. Since Nasdaq and AMEX stocks have a tilt towards smaller stocks, using the NAN universe to construct the size breakpoint will results into a lower median. We indeed show that the median log market capitalization is always higher under the NYSE criteria relative to the NAN criteria. More importantly, the market share of small stocks decrease when NAN breakpoints are used. Lastly, using the NAN breakpoints ensures that exactly 50% of the stocks are either small or long. Whereas, using NYSE breakpoints, approximately 80% of the stocks are in the small segments (assuming that microcaps are included). Therefore, using NAN breakpoints is likely to provide an overweight towards micro- and small-cap stocks.

2.3.3 Microcaps:

Closely related to the breakpoints (universe) is the inclusion and exclusion of microcaps as a construction choice. Microcaps are defined as stocks that are smaller than the 20th percentile of market equity for NYSE stocks. Fama and French (2008) find that microcaps accounts for 60% for the number of stocks, but only capture 3% of the total market capitalization. In addition, microcaps have the highest cross-sectional volatility of returns and large dispersion in sorting characteristics. From a practical perspective, these tiny stocks are out of reach for many (institutional) investors. In addition, microcaps are more expensive to short due to high shorting fees (Drechsler & Drechsler, 2014), may be illiquid, and have high transaction costs (Novy-Marx & 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, and using NYSE breakpoints. Figure 3 shows the effect of excluding microcaps on the size dimension of the CRSP universe. Excluding microcaps increases the median market capitalization. Furthermore, the market share of stocks below the median increases. Lastly, the number of stocks in the small segment decreases under the NYSE breakpoint when microcaps are excluded.

2.3.4 Financials:

Excluding financial firms from the sample is common practice in many empirical studies. The argument for this exclusion criteria is that financial services are fundamentally "different". 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. Most of the literature seem to follow Fama and French (1992) when it comes to excluding financial firms, with some studies including financials,such as (Stambaugh & Yuan, 2017). However, Fama and French (2015) and Fama and French (2018) do not (even implicitly) mention whether they include or exclude financial firms at all. Furthermore, including financials may impact factor returns, and some factors more than others. Especially, when factors are not hedged against industry exposure, financial companies may be overweighted or underweighted.

2.3.5 Industry Hedging:

Following up on the previous choice, we also consider industry hedging as a construction choice. The unconditional predictive power of stock characteristics may either stem from their across industries component or from their firm-specific (within industries component), or both (Ehsani, Harvey, & Li, 2021). A consequence of unconditional sorting is that factor portfolios obtain excess exposure (or too little) towards specific industries. To illustrate, constructing the unconditional value factor disproportionately overweights in sectors that contain stocks with high book-to-market ratios, such as utilities in the long leg. Whereas, the short value leg has excess exposure towards technology stocks.

Daniel et al. (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. Hedging these exposures is a choice that can be made in order to improve risk-adjusted returns[2]. A comparison of the standard and

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

industry-hedged factors shows that industry adjustment often improves factor performance, such as Asness, Porter, and Stevens (2000) and 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:

$$S^*_{i,t} = (S_{i,t} - \bar{S_{i,j,t}})/(S_{max,j,t} - S_{min,j,t}) \tag{9}$$

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

2.3.6 Independent versus dependent:

Independent sorting is the most commonly used sorting procedure deployed in the literature. A major drawback is that independent sorting may results in sparse portfolios, with the consequence that factor portfolio isn't well-diversified. In some cases, independent sorting may even result in empty portfolios. Especially in international samples or when data is scarce, this might be an issue. Dependent sorting alleviates the problem of sparse portfolios by sequentially stratifying stocks into portfolios. However, implementing a dependent sorting procedure raises the question 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.

2.3.7 Which size? June or most recent:

The 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 et al. (2020) do so when constructing the PEAD factor, and Ang et al. (2006) for the idiosyncratic volatility anomaly. For both choices there is no clear argument on why to use that choice. One argument in favour for the most recent market capitalization is that we use timely information to construct the size sort. On the other hand, this may result into more turnover, since we rebalance the size sorts each month, instead of each year.

2.3.8 Value-weighting vs. Equal-weighting:

There are several weighting schemes that a researcher can select when constructing a portfolio, and each will result in different properties for otherwise identical portfolios. The literature seems to focus predominantly on value- or equal weighting portfolios. Differences in the choice regarding weights result into a different the portfolio composition, and hence also in portfolio characteristics and performance. When using the value-weighting approach these exposures depend on the size of the specific companies. The risk and return will be driven predominantly by the largest companies in the investment universe unlike the equal-weight approach where each stocks contribution to the portfolio return is the same. From an industry perspective, value-weighted portfolios typically serve as a benchmark against which portfolio managers are evaluated, indicating the relevance of value-weighting. Nevertheless, many studies use equal-weighting when constructing factor portfolios (Hou et al., 2020). As a consequence, such equal-weighting results into the overweighting small- and microcaps in portfolio sorts, which in turn results in higher returns.

3 Model selection

3.1 Maximum Sharpe ratio

The ability of an asset pricing models to price assets depends on the extent in which its factors spans the mean-variance efficient portfolio. In case 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. Barillas and Shanken (2017) introduce 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. Why is that? Gibbons, Ross, and Shanken (1989) show that the gain of adding test assets to a factor model can be written as:

$$Sh^2(f, \Omega) - Sh^2(f) = \alpha' \Sigma^{-1} \alpha \tag{10}$$

$Sh^2(f, \Omega)$ denotes the maximum squared Sharpe ratio obtained from the factors f and assets Ω, and likewise $Sh^2(f)$ 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. The aim is to minimize equation 10, which implies that we minimize the mispricing that an asset pricing model creates. 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 formula 1 corresponds to maximizing $Sh^2(f)$. Hence, model selection can be done by comparing the maximum squared Sharpe ratio between models. Detzel et al. (2021) show that when (transaction) costs are ignored, model comparison based on squared Sharpe ratios favour models with high gross performance, even when trading costs are high. Hence, we consider both gross - and net factor returns, and report the maximum Sharpe ratio on an annualized basis.

First, we examine how specific construction choices, in isolation, affect maximum Sharpe ratio estimates. Figure 4 and Figure 5 shows the annualized maximum Sharpe ratio, gross and net respectively, by construction choice, averaged over factor models. First, we vary the breakpoints that are used to classify high and low characteristics. We either use the 30th-70th percentile (dashed bar) or the 20th-80th percentile (white bar). The former case yields an average annualized Sharpe of 0.64, 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 as we also see in our results. Second, including financials increases the Sharpe ratio on average from 0.63 to 0.66. Our previous results show that the size depth is not significantly altered when we include financials, however it allows for more factor breadth. Hence, this allows to make spread portfolios with a larger differences in the sorting characteristic. Third, the Sharpe ratio for independent and dependent sorts are approximately similar. Fourth, using the most recent market cap to construct increases the Sharpe ratio from 0.63 to 0.66 relative to using the market cap in June. Fifth, excluding microcaps improve the Sharpe from 0.59 to 0.70. Furthermore, eliminating industry exposures from factor returns increases Sharpe ratios, which is in line with Daniel et al. (2020). However, taking transaction costs under account, industry neutralization leads to a lower net sharpe ratios compared to unhedged factor portfolios. In addition, using NAN breakpoints improves gross Sharpe ratios from 0.54 to 0.75, which is the largest increases within our set of choices. However, the impact is small on a net-basis. By using NAN-breakpoints, we also allow more small caps to enter the portfolio construction, which have larger transaction costs. Lastly, equal weighing portfolios improves the gross Sharpe ratio on average compared to value-weighting portfolios from 0.58 to 0.71. On the other hand, equal-weighted factor portfolios are tilted towards small caps with higher transaction costs. Hence, the performance on a net-of-cost basis is lower than for value-weighted factor portfolios. These results imply that construction choices can materially affect factor performance.

How does this implication translate in model selection exercises? Recent literature seems to compare factors using their "original" construction method, thereby comparing factors without taking differences in construction method into account. We explicitly construct each possible construction method and compare factors on an "apples-to-apples" basis. Figure 6a reports the average maximum Sharpe ratio of a factor model (using value-weighted return) whereby we average across all possible construction methodologies. Around the average, we also plot the 2σ-spread of the Sharpe ratio of a factor model. Following Detzel et al. (2021) we separately run this model selection exercise for gross and net returns[3]. The gross sharpe ratio for the FF5 model, averaged over construction methods, is 1.13. Replacing operating profitability with cash profitability increases this to 1.37. Adding the momentum factor further improves the gross Sharpe ratio to 1.50. The Q4, BS6, and DHS factor model have a gross Sharpe ratio of 1.39, 1.73, and 1.70, on average. Note that we compare factor models using the same construction choices, and then average over this set of choices. By doing so, we can compare factor models on a "apple-to-apple" basis, keeping construction choices constant. If a researcher would've chosen different construction methods for different factor models, there would be a possibility that model rankings differ, indicated by the error bars. Suppose, that a researcher picks 80-20 breakpoints for the PEAD factor, but 70-30 for factors in the BS6 model, then it could be the case that the DHS achieves a higher gross Sharpe ratio than the BS6 model. The gross Sharpe ratios, however, do not represent what is actually achievable by investors. The model ranking for net factor models look slightly different. FF6, with cash profitability, has a net Sharpe ratio of 0.87, whereas the BS6 model has a net Sharpe ratio of 0.82. The DHS model has the largest net Sharpe ratio, which equals 0.92. In addition, the DHS model also has the lowest standard deviation in net Sharpe ratios, which equals 0.07. Hence across construction choices, the DHS robustly achieves a "high" net Sharpe ratio. The BS6 model has the largest standard deviation, equal to 0.19, indicating that the net Sharpe ratio fluctuates across construction methods. Figure 6b shows the model selection results, when we equal weight factor returns. Compared to the value-weighted results, average gross Sharpe ratios increases, but also the Sharpe standard deviation. The average net Sharpe for BS6 is slightly larger than that of the DHS model, but the BS6 has a very wide Sharpe standard deviation, indicating that the performance hinges strongly on the construction choice.

[3] For mean-variance analysis with transaction costs, we follow the approach in Novy-Marx and Velikov (2016): We estimate mean-variance optimal weights by using a long and short versions of all the assets in the portfolio, net of transaction costs, subject to a no-shorting constraint on portfolio weights.

Table 3 reports the portfolio weights that correspond to the ex post mean-variance efficient portfolios constructing from the candidate factor models, where we average the weights across all construction methodologies. Between brackets, we report the standard deviation in weights, which we obtain from all our 256 construction methods. The upper panel provides the weights allocated in the portfolio for a mean-variance investor, given that there are no transaction costs. Since the factors are constructed on an equal-basis, the weights can be compared directly. Panel A shows a large discrepancy in optimal weights within factor models. The Fama-French 5 factor model allocates 48.7% weight towards CMA on average. However, a researcher that randomly picks a construction choice, may find that the weight on CMA varies largely between 28.5% and 68.9% for a two standard deviation change. HML has a small average weight of 1.9% in the 5-factor model, however for some construction methods the HML might be negative (-16.7% for a two standard deviation decrease) or highly positive (20.5% for a two standard deviation increase). The momentum-augmented Fama-French model and cash-based Fama-French model allocate 13.4% and 36.7% towards UMD, and RMW(cp) respectively. Both, however, show large variation and may imply that the factors have a relative low, or high, important for some construction versions. The Q4 model tends to improve on the 5-factor model by swapping out on the investment factors for their ROE factor, which use more timely information (quarterly ROE data). Compared to the Fama-French models, the Q factor seems to have more stable weights. For example, the I/A factor ranges between 32.1% and 48.1% (given a two standard deviation interval). Similarly, the BS6 models improves 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 20.5%, with a standard deviation of 4.6 across construction methods. The monthly updated value factor receives a relatively larger weight (compared to FF-models) of 28.7%, with a standard deviation of 6.9. Lastly, the DHS models allocates 22.9% and 58.9 to the FIN and PEAD factor, respectively.

Panel B shows the weights of the mean-variance efficient portfolios using net factor returns. To reiterate, 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 weight to the market increases, relative to gross returns. Since, transaction costs are incurred, factors are less profitable, hence more weight is allocated towards the markets. Most factor weights decreases due to the transaction costs. CMA, for example, in the FF5 model decreases from 48.7% (gross) to 26.5%. In addition, due to the no-shorting constraint, zero weights are allocated to factors with high transaction costs and negative net alpha. One example of such a case is the PEAD factor. It has a net weight of 8.3%, compared to 58.9% gross weight, and a 12.8% standard deviation. For multiple construction choices, PEAD has a negative net alpha, thereby binding the no-short constraint and consequently received zero weight. The net DHS model predominantly consists of the financing factor (52.9%) and the market factor (38.8%), on average.

3.2 Frontier expansion

The results from the previous section indicate that model performance and its underlying weights depend on construction methods. Next, we aim to measure the extent to which adding factors of a model "M1" to those of model "M0" expands the efficient frontier. To this end, we implement the multi-factor version of the generalized alpha of Novy-Marx and Velikov (2016): 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 M0:

$$MVP_{M1 \cup M0,t} = \alpha + \beta MVP_{M0,t} + \epsilon_t \tag{11}$$

We present the results in table 4, where we show the results of these spanning regression for each pair of models using gross returns. The results, based on gross returns, show that most of the models expand the efficient frontier when added to others, on average. Adding the BS6 factors or DHS factors to FF models greatly improves the efficient frontier with alphas between 0.08 and 0.48 per month. However, across construction methods, we find large standard deviation in estimated alphas. For some, construction methods, the estimated alpha drops to zero. For example, the adding the FF5 model to the Q4 model expands the efficient frontier on average, with an estimated average alpha of 0.06 per month, with a standard deviation of 0.05. Under some construction methods, the estimated alpha is lower, and closer to zero. Table 4 panel B shows the results when we focus on factor returns net of transaction costs. Estimated alphas are closer to zero due to transaction costs. Adding BS6 factors to FF models expand the efficient frontier between 0.05% and 0.07% per month, with a standard deviation of 0.06. Hence, there are construction methods for which the added value of the BS6 factors to the FF models is zero. The DHS improves FF models between 0.14% and 0.18% per month with standard deviations between 0.06% and 0.07%. Hence, there are less construction methods that reach alphas closer to zero for the DHS model compared to the BS6 model. Again, our results imply that construction methods may influence model selection exercises as shown by the large alpha standard deviations.

3.3 Economic Significance

We quantify the economic significance, in table 5, by reporting how much percent the maximum Sharpe ratio would be increase if we would add the additional factors to the base model "M0" for each pair. This directly reflects the gain that could be realized by a mean-variance investor. Panel A shows the results for gross returns. In most cases, adding one model to a base model improves the Sharpe ratio of the combined model. On a gross basis, adding the Q4 factors to Fama French models improves the Sharpe ratio between 7.3% (FF6(cp) as base) and 29.6% (FF5 as base). Similarly, adding the BS6 model or the DHS model to Fama-French models also greatly improves the Sharpe ratio. This improvement, however, depends on the construction choice. Between parentheses, we also report the standard deviation of improvement in Sharpe ratios, across construction methods. For example, on average, the BS6 model improves the FF6(cp) model by 48.4%, but also has a standard deviation of 21.5%. This implies that there are construction methods for which the improvement is less (a two standard deviation decrease would imply a mere improvement of 5.4%). In panel B, we present the results using net returns. This provides a more realistic view of the extent in which the investment opportunity set improves when adding factors. Adding the Q4 factors, in this case, improves the Fama-French models between 4.5% (FF6(cp)) and 9.2% with standard deviations between 6.1% and 8.8%. For quite some construction methods, the Q4 factor adds little to no improvement relative to the FF models. Likewise, the BS6 and DHS also improve FF factor models less compared to gross returns, since these models contain factors with relative high turnover and transaction costs. For example, adding the BS6 factors to FF5 improves the Sharpe ratio by 16.1% (for gross returns this was 84.4%) on average, with a standard deviation of 15.8%. In many instances, the BS6 improves the FF5 model little. The main takeaway here is that the improvement in Sharpe ratio, when adding additional factors, is not only a function of expected returns, variances, and correlations of and among factors, but also factor construction choices.

3.4 In-sample and Out-of-sample estimation

The main criteria that we use to select models is the maximum Sharpe ratio. We use full-sample estimates to calculate this metric. 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 estimate of the true weights. Consequently, the estimates of the maximum Sharpe ratio are biased upwards. This bias also becomes larger in smaller samples, since the parameter estimates have more sampling error. Note, that the bias in the estimates of the maximum Sharpe ratio is especially problematic for comparing non-nested models, such as the Q-factor model vs. 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, hence it allows to test whether two models have statistically different squared Sharpe ratios. In addition, the advantage of the bootstrap approach allows to test how often one model outperforms the other, or vice versa.

The bootstrap procedure is as follow: we 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. Note that a month might appear multiple times in the IS sample if the pair is drawn multiple times. 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, the OS Sharpe ratios are free of such a bias, since monthly returns are approximately serially uncorrelated. For each construction rule r we run 100.000 simulation runs. For each run, we compare the maximum Sharpe ratio between models, and count how many times a model has a higher max Sharpe ratio than an other model. By doing so, we can calculate the both the IS and OS 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 6 shows the winning-probability estimates obtained from the bootstrap simulations. Panel A shows the in-sample estimates, which should be interpreted with caution, since the in-sample Sharpe ratios are upward biased and based on 300 observation months. We find that the FF6(cp) model outperforms the Q factor model in 57.2% of the sample, and its Sharpe ratio (1.65) is slightly higher than that of the Q factor (1.61). The BS6 model outperforms most of the models, with win-probabilities over 63%. It is, on average, the model with the highest Sharpe ratio in 63.6% of all simulation run. However, the standard deviation is 26.41%, implying large variation across construction methods. The DHS model is the second-best model in this simulation with an average Sharpe ratio of 1.85 and is the model with the largest Sharpe ratio in 32.9% of all simulation runs Panel B presents the out-of-sample (OS) results. The results are qualitatively similar to the in-sample results: the BS6 remains the model with the highest probability to have the largest Sharpe ratio in the simulation runs, with an average probability of 52.4% and a variation of 29.43%. Panel C considers the in-sample estimates using net returns. The 6-factor model with cash profitability is the best model in 13.6%, compared to 3.5% when using gross settings. Taking transaction costs into account, the BS6 model is no longer the model with the highest win-probability (35.1%). The DHS model has a win-probability of 45.6%. Panel D shows the results for the out-of-sample estimates using net returns. Similar to panel C, the DHS remains the model with the largest winning-probability (53.1%). In addition, the variation around the Sharpe ratio for the DHS model is also relative low (0.12) compared to all other models (0.19 and above).

4 Factor composition across construction choices

We have shown that factor returns vary significantly across different sets of construction choices. In this section, we study how variation in construction choices affect fundamental portfolio characteristics that, in turn, drive portfolio performance. We consider the factor exposure of a portfolio, its (il)liquidity, and to what extent a factor portfolio is diversified.

4.1 Factor exposure

The expected return of a well-diversified factor portfolio is directly related on 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. Do construction choices have an impact on the factor exposure of factors? In table 8 we examine the impact of construction choices on factor exposure. We define factor exposure by creating a normalized factor score. Equation 12 shows that 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.[4]

$$Normalized\ factor\ score_{i,v,t} = \frac{Variable_{i,v,t} - Mean_{v,t}}{Max_{v,t} - Min_{v,t}} \quad (12)$$

In both the long, and the short side of the long-short portfolio, we aggregate the normalized factor scores to 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. To examine the impact of construction choices on the factor portfolio, we run a fixed effect regression in which we include factor- and time fixed effects. This regression is run over the 256 construction choices and the 10 factors we defined in section 2.[5]

Table 8 shows the results of the fixed effect regressions. We find that out of the 8 choices, 5 have a significant (at the 5% significance level) impact on the factor exposures. Including 30% and 70% breakpoints has a significantly negative impact, implying that using 20% and 80% breakpoints for the variable sorts result in more factor exposure. If variable scores are linearly related to expected returns, this would be expected. Including the financial industry shows an insignificant impact on factor exposures, therefore the impact of in- or excluding financial stocks inconclusive. The impact of independent or dependent portfolio construction and using the market capitalization of last June as opposed to last months' have insignificant coefficients as well.

Including micro-cap stocks, as defined by the stocks with the 20% smallest market capitalization, has a significant and positive coefficient. Including these stocks adds breadth to the universe and adds positively to the factor exposure of the long-short portfolios as opposed to excluding them. Using NYSE breakpoints and using a value-weighted weighting both show a significant negative coefficient. A consequence of both nan-breakpoints and an equal-weighted weighing scheme as well as including micro capitalization stocks is that stocks with a smaller market capitalization get a larger weight in the portfolios all else equal.

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

[5] As the FIN factor uses two variables in its' construction, we have 11 variables, However, we adjust the weight to 0.5 for these two variables to arrive at a weighing of 1 for every factor, such that every factor has a weighting of 1.

4.2 Liquidity

In addition to the impact on factor exposure of the various construction choices, construction choices may have impact on the liquidity of a portfolio as well. Stocks with low liquidity, such as microcaps, may have high transaction costs, and other frictions, such as bid-ask spreads, and hence directly impact both the gross and net returns of factor portfolios. We measure the liquidity of the portfolios by aggregating stock illiquidity to portfolio illiquidity following Amihud (2002). Stock illiquidity is defined in equation 13. We measure stock illiquidity as the ratio of the daily absolute return to the dollar trading volume on that day, $|R_{imd}|/VOL * P_{imd}$. $|R_{imd}|$ is the absolute return on stock i on day d of month m and $VOL_{imd} * P$ is the dollar trading volume for stock i on day d of month m to which we refer as $VOLD_{imd}$. We aggregate the daily illiquidity measure to a monthly ratio by averaging over the days for which data was available in said month. As it is an measure of illiquidity, a higher (lower) value implies lower (higher) liquidity:

$$ILLIQ_{im} = \frac{1}{D_{im}} \sum_{t=1}^{D_{im}} \frac{|R_{imd}|}{VOLD_{imd}} \tag{13}$$

For the 10 factors used throughout this paper, we have computed the illiquidity measure of each of the 256 construction choices. We end up with 2,560 portfolios, containing 600 illiquidity-month observations. To test the impact of construction choices on the factor portfolio, we run a fixed effect regression in which we include factor- and time fixed effects. Table 7 shows the results of the fixed effect regressions. The results show that 30th and 70th percentile breakpoints as opposed to 20th and 80th percentile breakpoints significantly improve the liquidity of the factor portfolio. In addition, independent portfolio sorting as opposed to dependent portfolio sorting, imposing industry neutrality, the use of NYSE breakpoints and using a value-weighted weighting scheme ceteris paribus significantly improve the liquidity of the factor portfolio as well. Using the market capitalization of last June in sorting on size and including micro-caps ceteris paribus have a significant negative impact on portfolio liquidity. The inclusion of financial stocks has an insignificant impact on portfolio liquidity.

4.3 Risk diversification

The construction choices also affect the portfolio risk. Suppose that we have selected N assets $(X_1, ..., X_N)$ for a factor portfolio with weights $(\omega_1, ..., \omega_N)$, stock variances $(\sigma_1, ..., \sigma_N)$, and covariances σ_{ij} (with $i \neq j$), then the portfolio risk is defined as:

$$\sigma_p^2 = \underbrace{\sum_{i=1}^{N} \omega_i^2 \sigma_i^2}_{(1)} + \underbrace{\sum_{i=1}^{N} \sum_{i \neq j}^{N} \omega_i \omega_j \sigma_{ij}}_{(2)} \tag{14}$$

The first term (1) is the contribution of individual stock variances to the total portfolio risk. The second term (2) is the contribution of covariances among stocks to the total portfolio risk. For large N the first term tends to be very small, and the second term converges to an "average covariance". Factor portfolios with few stocks may have a higher variance term, resulting into under-diversification and lower risk-adjusted return. For example, if we use 20th-80th breakpoints instead of 30th-70th we select less stocks into the factor portfolio, thereby introducing idiosyncratic or unpriced risk. In addition, selecting stocks that are highly correlated increases the second term, especially when the portfolio consists of a few stocks. Another example is that unhedged factors may tilt towards certain industries, hence selecting stocks in the same industry with higher correlations. Choosing to hedge out industry exposure will reduce correlations among assets, and decrease the second term. These examples illustrate that construction choices directly impact portfolio risk diversification, and hence Sharpe ratios.

To quantify risk diversification we use the diversification ratio (DR) of Choueifaty and Coignard (2008), which is defined as the ratio of a portfolio's weighted average of individual volatilities over its portfolio volatility:

$$DR = \sum_{i=1}^{N} \Big(\omega_i \sigma_i \Big) / \sigma_p \tag{15}$$

Essentially, the DR measures to what extent diversification is gained from holding assets that are not perfectly correlated. When all stocks are perfectly correlated, then DR equals 1. We estimate DR by using 36-month rolling volatilities for all (versions of) factors. We regress the estimated DRs on construction choices dummies using a panel regression with factor- and time fixed effects. The results are shown in Table **XX**.

5 What about 2x3x3 sorts?

The previous results focus on 2x3 sorts à la Fama and French (1993). An important and popular construction choice that we have not included in our analysis is the 2x3x3 sorts. The IA and ROE factors from Hou et al. (2015), for example, are constructed using the 2x3x3 independent sorting procedure. They independently sort on market capitalization, annual change in total assets, and quarterly return on equity. Why don't we include the 2x3x3 construction choice?

First, there is no clear guidance on which additional sorting one should pick to include in the 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, and which additional characteristic will 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 pick from.

Second, using the 2x3x3 sorting methodology may lead to sparse or even empty portfolios in some instances. We construct an additional 256 versions of each factor, using 2x3x3 sorting instead of 2x3 sorting, with the HML factor as the second sorting characteristic when we construct 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 9, we count for how many months we obtain empty portfolios for these factors. When we construct 2x3x3 factors using 30-70 breakpoints, we find that most factors have no empty portfolio. The only except is RMW and FIN, with 0.1 months and 4.5 months (out of 600) of missing data on average. Using 80-20 breakpoints limits the cross-section, and allows the occurrence of empty portfolios to increase. For RMW, 24 months are missing on average, and for the FIN factor even 82 months. Panel C and D consider the NYSE and NAN breakpoints, respectively. We again find around missing months for RMW and FIN, in particular. Using the NAN breakpoints creates a wider universe to select stocks from, hence empty portfolios occur less often compared to NYSE breakpoints.

Moreover, 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. In table 11 we show how many positions are taken, on average, in each portfolio used to construct the long and short leg[6]. We find that the extreme portfolios typically contain below 50 stocks. For example, Big-High-Low ("BHL") contains an average of 49 stocks when excluding microcaps, and 38 when we use NYSE breakpoints. Big-Low-High contains between 34 and 74 stocks, across construction choices. Especially during the early periods of the sample, the number of stocks are even less than mentioned in the table.

Next, we compare the impact of 2x3x3 sorting of the factor performance of the Q-factors. In figure 7 we show the Sharpe ratio variation of the Q factors for the 2x3 and 2x3x3 sorts, using gross returns. We find that the 2x3x3 sorting methodology tends to increase the distribution of Sharpe ratios higher relative to the 2x3 sorting methodology. For the 2x3 IA factor, we find a median Sharpe ratio of 0.72, whereas the median for the 2x3x3 IA factor equals 0.91. For the 2x3 and 2x3x3 ROE factor, we find Sharpe ratios of 0.71 and 0.80, respectively. We also find that the range of the Sharpe ratio widens when the 2x3x3 sorting procedure is implemented relative to the 2x3 sort. The minimum Sharpe ratio for the 2x3 and 2x3x3 IA factors equal 0.20 and 0.36, respectively. For the ROE factor this equals, 0.42 and 0.47. Likewise, the maximum Sharpe ratio for the 2x3 and 2x3x3 IA factors equal 1.31 and 1.64, respectively. For the ROE factor this equals, 1.03 and 1.13.

Lastly, we consider the impact of 2x3x3 jointly, with all other construction choices, on the Sharpe ratios of the Q factors. These results are shown in figure 8. 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 0.84, whereas 2x3x3 yields a Sharpe ratio of 0.99. Hence, the 2x3x3 sorting methodology is a construction choice that is able to consistently increase the risk-adjusted return of factors.

6 Conclusion

Within empirical asset pricing, character-based sorting is perhaps the most popular way to construct factors. This paper stresses out that constructing factors involves a large number of choices, leading to "degrees of freedom" to researchers. Especially since there is no consensus on construction methods, the degrees of freedom involved may arguably allow for p-hacking: researchers could pick construction choices such that the resulting factor meet certain statistical and performance-related hurdles, such as high t-statistics, and Sharpe ra-

[6] For brevity, we do not show portfolios that are not used to construct the I/A and ROE factor, such as the big-neutral-neutral portfolio.

tios.

Construction choices indeed impact factor returns. Using over 250 different combinations of construction choices, we show large and significant variation in Sharpe ratios in factor returns. Such variation also materially impacts model selection exercises when comparing models. Mean-variance weights vary tremendously across construction methods, with some factors received zero weight when taking transaction costs into account. Maximum Sharpe ratio of factor models also show wide variation across construction methods. In addition, our bootstrapping approach shows that the probability of having the highest Sharpe ratio has standard deviations of up to 41%. Next, we study the efficiency of factor portfolios, and point out important consequences of using specific construction choices. We attribute differences in factor performance across methods to differences in factor and liquidity exposures.

Our study has two important implications for researchers. First, our results indicate that factor returns depend on factor construction choices. Hence, a researcher should always run a series of robustness tests or consider a multiple construction methods. For example, a recommended common practice for the future is to plot the distribution of Sharpe ratios of a new factor for a set of construction methods. Second, we show the impact of equalizing construction methods across factors in model selection exercises. Factors should not be compared against each other when their construction method differs, but should always be compared on an equal basis, i.e. using the same construction choices. Hence, researchers should always 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.

Choueifaty, Y., & Coignard, Y. (2008). Toward maximum diversification. *The Journal of Portfolio Management*, *35*(1), 40–51.

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 et al., K. (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*.

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., & 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.

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.

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

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.

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.

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.

7 Tables and Figures

Figure 1: **Gross Sharpe ratio variation within factors:** This figure plots the distribution of annualized gross Sharpe ratios for long-short factor returns, 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 2. The sample, to calculate these Sharpe ratios, runs from January 1972 till December 2021.

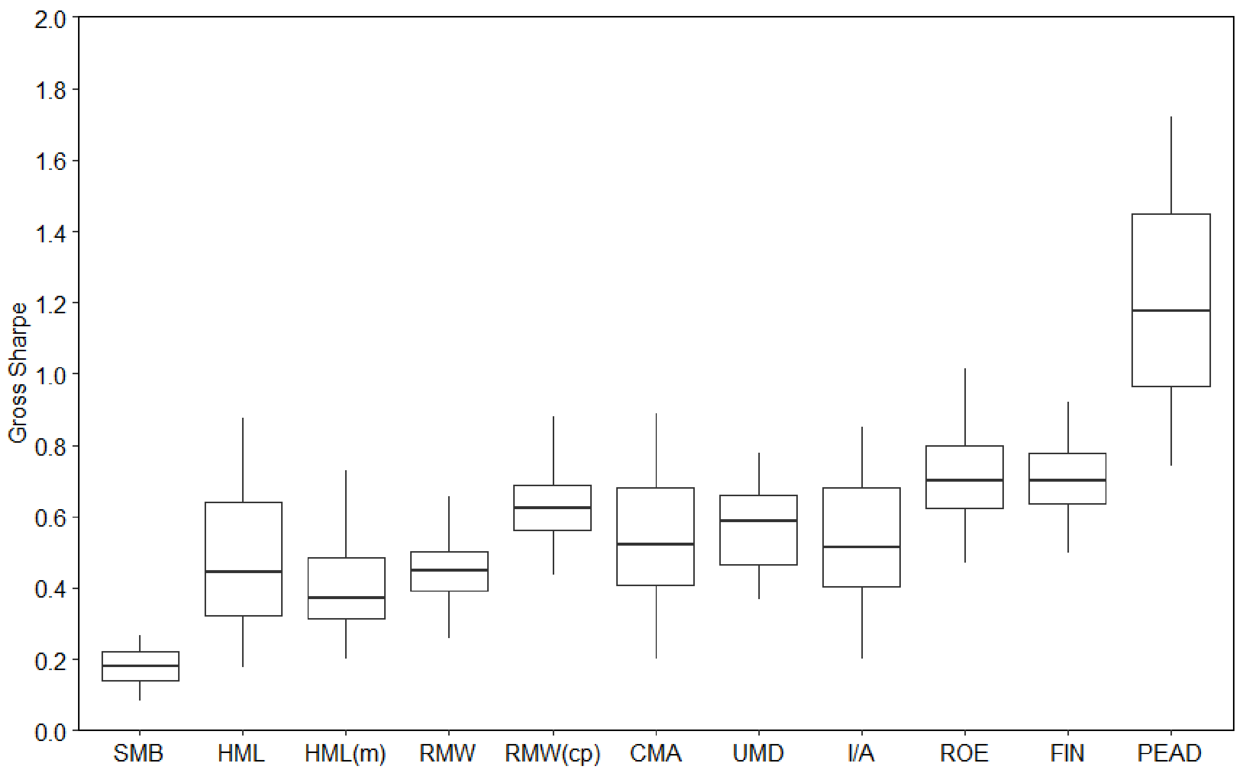


Figure 2: **Net Sharpe ratio variation within factors:** This figure plots the distribution of annualized gross Sharpe ratios for long-short factor returns, where a factor is constructed N times by using the N different factor construction methods. The black solid line within the box plot shows the median Sharpe ratio. The upper (lower) bound shows the 75th (25th) percentile. The factors and their definitions are from table 2. The sample, to calculate these Sharpe ratios, runs from January 1972 till December 2021.

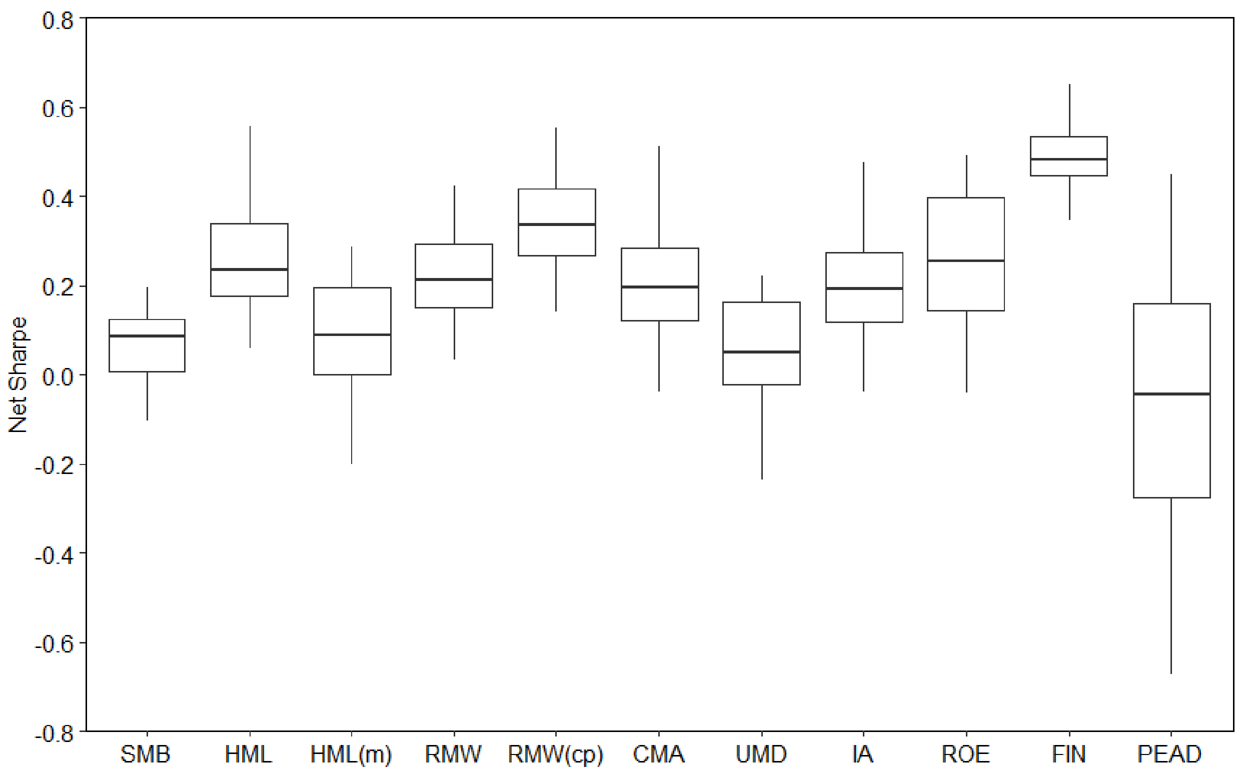


Figure 3: **Sample distribution size statistics conditional on exclusion criteria:** The figures show from top to bottom 1) the Median breakpoint of market capitalization which is used to assign stocks to either one of the Small or one of the Big portfolios 2) the % of market capitalization assigned to the Small portfolio and 3) the % of stocks assigned to the Small portfolios. The left-hand side uses NYSE breakpoints, the right-hand side uses NAN breakpoints. From left to right in the graph, the figures show the values for 1) no exclusion criteria, 2) excluding Micro cap stocks (20% smallest stocks), 3) excluding Financial stocks, and 4) both excluding the smallest 20% stocks and excluding Financial stocks.

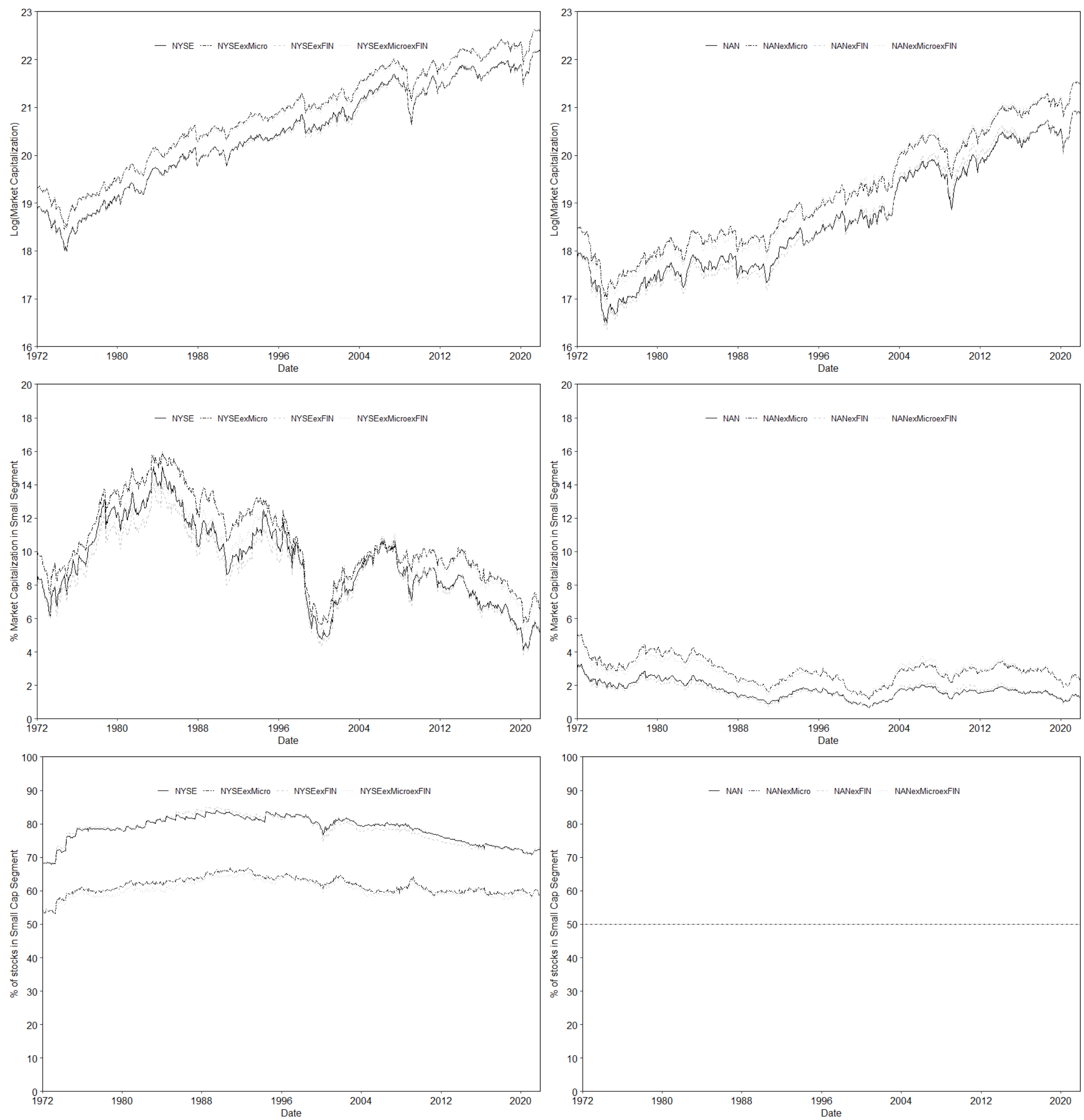


Figure 4: **Construction choices and gross squared Sharpe ratios:** This figure shows the impact of construction choices on the squared Sharpe ratio averaged over factors. Sharpe ratios are computed on a gross-basis and are annualized. We consider eight choices. "30-70" refers to the use of the 30th and 70th percentile as breakpoints in the sorting procedure ("Yes") or the use of the 20th and 80th percentile ("No"). "Financial" means that companies in the finance sector are included ("Yes") or excluded ("No") in the factor construction. "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"). "Micro" indicates whether we exclude stocks with the smallest 20% market cap based ("Yes") or not ("No"). "Neutral" means that portfolio sorts are constructed using industry-adjusted characteristics ("Yes") or the standard characteristics ("No"). NYSE indicates whether the NYSE cross-section is used to construct breakpoints ("Yes") or the full NYSE-AMEX-Nasdaq cross-section ("No"). "VW" indicates whether factors are calculated by using value-weighting ("Yes") or equal-weighting ("No"). Monthly factor returns are constructed using data spanning January 1972 and December 2021.

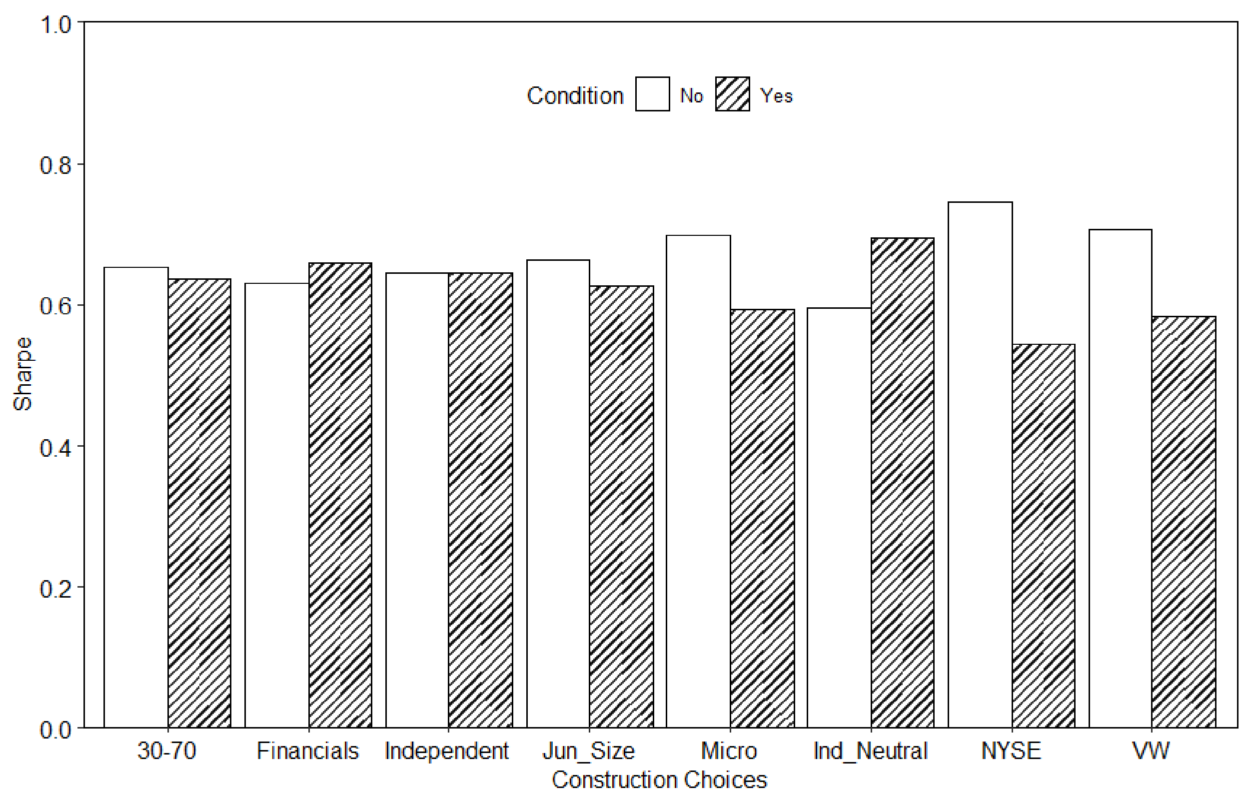


Figure 5: **Construction choices and average net Sharpe ratios:** This figure shows the impact of construction choices on the Sharpe ratio averaged over factors. Sharpe ratios are computed on a net-basis and are annualized. We consider eight choices. "Financial" means that companies in the finance sector are included ("yes") or excluded ("no") in the factor construction. "Independent" refers to the use of independent sorting ("yes") or dependent sorting ("no"). "BP" 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"). "Neutral" means that portfolio sorts are constructed using industry-adjusted characteristics ("yes") or the standard characteristics ("no"). "Recent" indicates that we use the one-month lagged market capitalization ("yes") or the market capitalization of June ("no"). "VW" indicates whether factors are calculated by using value-weighting ("yes") or equal-weighting ("no"). Monthly factor returns are constructed using data spanning January 1972 and December 2021.

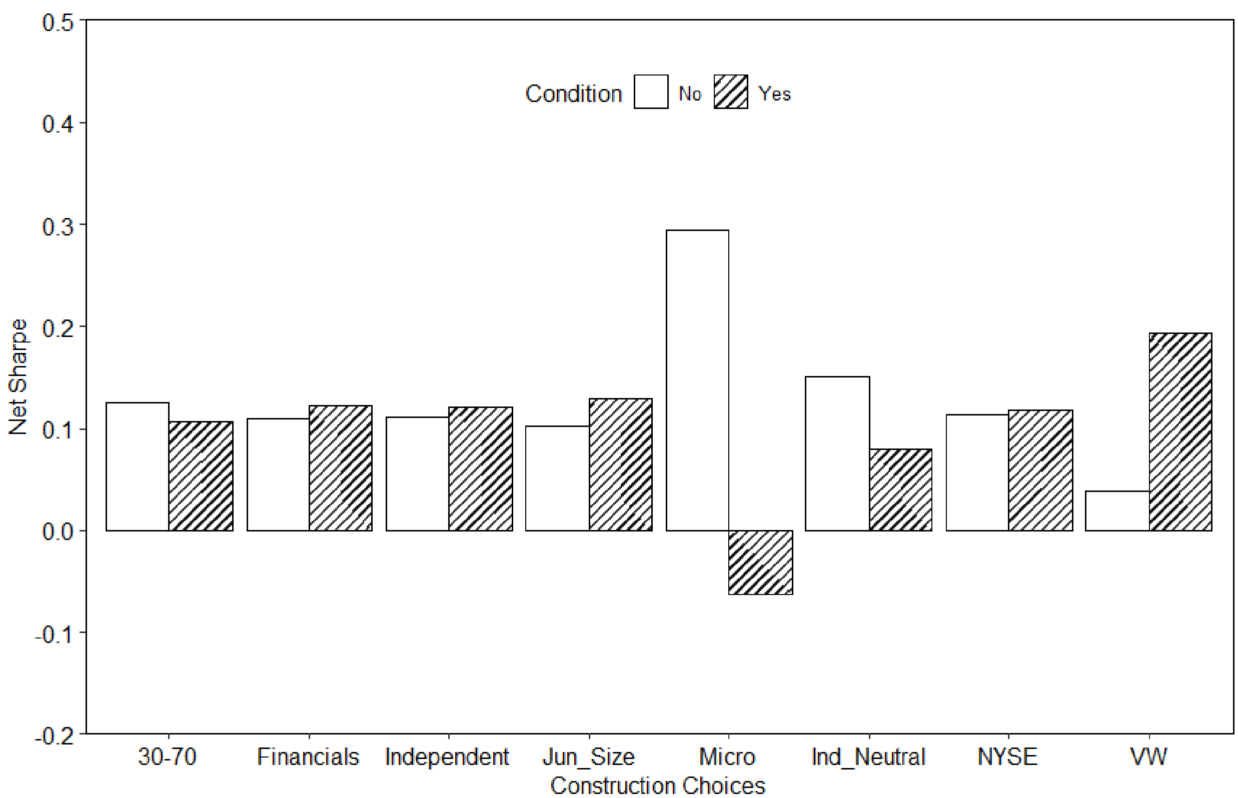
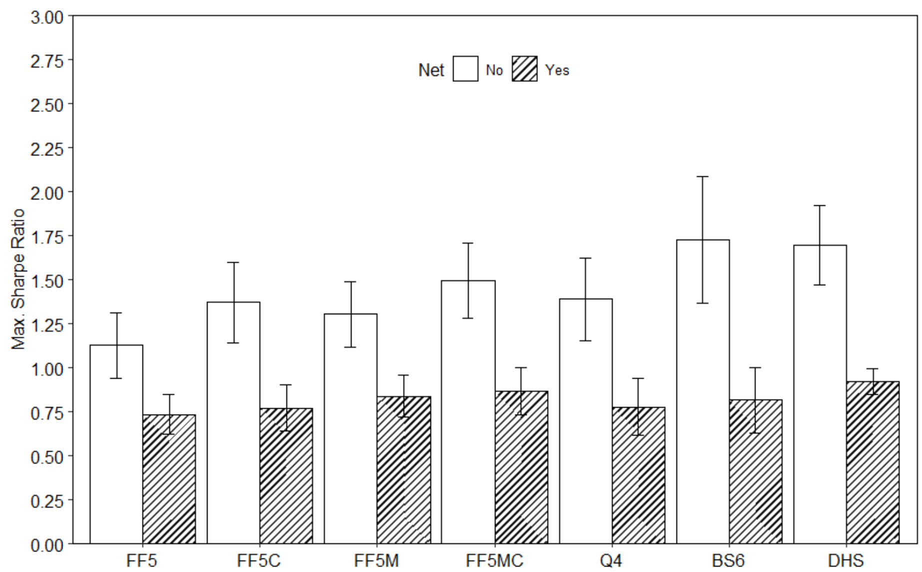


Figure 6: **Selecting factor models:** This figure shows the maximum gross Sharpe ratio (annualized) from the factors from the factor models listed on the horizontal axis. The white bar shows the maximum Sharpe ratio obtained by using gross factor returns. The dashed bar shows the maximum Sharpe ratio using net factor returns, taking transaction costs into account. The error plot shows the variation in the maximum Sharpe ratios for a given factor, across construction choices. The data runs from January 1972 till December 2021. Panel a shows the results for value-weighted portfolios, panel b shows the results for equal-weighted portfolios.

(a) Value-weighted portfolios



(b) Equal-weighted portfolios

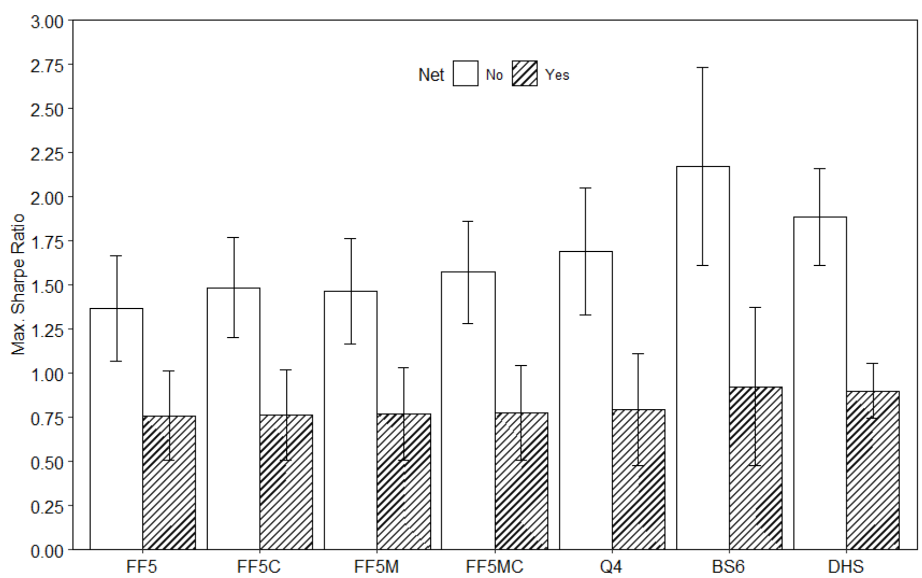


Table 1: **Summary Statistics:** This table reports the monthly average return (in %), standard deviation, t-statistics, and Sharpe ratio of the factors listed in table 2. Factor turnover is defined as the sum of the turnover of the long leg and short leg. In addition, we report these statistics using factor returns net of transaction costs. Lastly, we report the average and the median turnover and transaction cost per day. The data runs from January 1972 till December 2021.

	Value-Weighted						**Equal-Weighted**					
	Gross		**Net**		**TO**	**TC**	**Gross**		**Net**		**TO**	**TC**
	Avg	Sh	Avg	Sh	Avg	Avg	Avg	Sh	Avg	Sh	Avg	Avg
SMB	0.17	0.18	0.06	0.06	20.0	0.12	0.21	0.23	-0.03	-0.07	25.7	0.24
HML	0.39	0.48	0.22	0.27	26.8	0.16	0.58	0.72	0.23	0.29	34.9	0.34
HML(m)	0.42	0.41	0.10	0.09	53.1	0.32	0.78	0.74	0.23	0.22	57.7	0.55
RMW	0.35	0.45	0.18	0.22	28.4	0.17	0.26	0.31	-0.09	-0.12	36.9	0.36
RMW(cp)	0.40	0.64	0.21	0.34	31.6	0.18	0.37	0.54	-0.00	-0.02	40.1	0.37
CMA	0.31	0.55	0.12	0.21	34.6	0.19	0.51	0.93	0.12	0.22	42.4	0.39
UMD	0.71	0.57	0.08	0.05	117.3	0.64	0.57	0.44	-0.34	-0.26	109.6	0.91
IA	0.31	0.54	0.12	0.20	34.6	0.19	0.49	0.89	0.10	0.19	42.4	0.39
ROE	0.66	0.72	0.24	0.25	82.7	0.41	0.69	0.69	0.07	0.05	84.3	0.63
FIN	0.65	0.70	0.45	0.49	39.2	0.20	0.79	0.84	0.44	0.46	43.8	0.35
PEAD	0.57	1.19	-0.02	-0.07	120.9	0.59	0.63	1.44	-0.23	-0.53	115.2	0.86

Table 2: **Factor models:** This table lists the non-market factors used by the asset pricing model, indicated by a ✓in the column below the model name. Furthermore, we provide four properties of the factor construction methodology: the sorting characteristic ("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 (FF5M) denote the Fama and French (2015) five-factor model (augmented with UMD). Q4 denotes the Hou et al. (2015) four-factor q-model. BS6 denotes the Barillas and Shanken (2018) six-factor model. FF5$_c$ and FF5M$_c$ denote versions of the FF5 and FF5M, respectively, that use cash-based operating profitability instead of accruals operating profitability Fama and French (2018).

					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 3: **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 all possible construction methodologies. Within brackets, we show the standard deviation of the optimal weights that occur within our set of possible construction methods. Panel A reports weights obtained from gross factor returns. Panel B shows the weights using factor returns net of transaction costs. The sample period spans January 1972 through December 2021.

Panel A: Gross	Mkt	SMB	HML	RMW	CMA	UMD	RMW_{cp}	IA	ROE	HML_d	FIN	PEAD
FF5	19.0	6.4	1.9	23.9	48.7							
	(3.7)	(3.2)	(9.3)	(7.3)	(10.1)							
FF5M	17.4	8.2	8.0	18.7	34.3	13.4						
	(3.7)	(2.4)	(8.7)	(6.8)	(11.2)	(3.6)						
FF5C	17.3	9.4	0.6		35.9		36.7					
	(2.9)	(3.0)	(8.8)		(11.0)		(8.1)					
FF5MC	16.3	10.4	5.6		26.7	10.8	30.2					
	(3.0)	(2.4)	(8.9)		(10.9)	(3.1)	(7.0)					
Q4	16.7	13.0						40.1	30.2			
	(2.6)	(1.9)						(4.0)	(3.9)			
BS6	13.8	8.4				20.5		9.0	19.5	28.7		
	(3.5)	(2.9)				(4.6)		(6.9)	(4.5)	(6.9)		
DHS	18.2										22.9	58.9
	(2.7)										(6.2)	(7.9)
Panel B: Net	Mkt	SMB	HML	RMW	CMA	UMD	RMW_{cp}	IA	ROE	HML_d	FIN	PEAD
FF5	40.0	1.3	18.8	13.4	26.5							
	(22.9)	(2.4)	(14.7)	(13.3)	(19.0)							
FF5M	38.9	1.6	20.0	11.9	23.0	4.6						
	(23.5)	(2.6)	(14.1)	(12.0)	(17.5)	(4.8)						
FF5C	37.2	3.1	16.0		23.9		19.8					
	(23.8)	(3.8)	(13.9)		(18.9)		(16.9)					
FF5MC	36.6	3.4	16.8		22.0	18.0	3.3					
	(24.1)	(3.8)	(13.7)		(17.5)	(16.1)	(3.6)					
Q4	41.4	5.4						31.8	21.5			
	(29.7)	(4.9)						(17.2)	(10.9)			
BS6	39.6	4.0					4.6	21.0	19.8	11.0		
	(30.8)	(4.0)					(6.1)	(14.5)	(9.9)	(11.1)		
DHS	38.8										52.9	8.3
	(6.5)										(8.6)	(12.8)

Table 4: **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. Panel A reports the results using gross returns. Panel B reports the results using net returns, taking transaction costs into account. The data runs from January 1972 till December 2021.

Panel A: Gross				**Union Model (M1)**			
Base "M0"	FF5	FF5$_c$	FF6	FF6$_c$	Q4	BS6	DHS
FF5		0.10	0.11	0.17	0.24	0.48	0.35
		(4.24)	(4.02)	(5.89)	(6.74)	(12.09)	(9.70)
		[0.03]	[0.05]	[0.05]	[0.10]	[0.17]	[0.08]
FF5$_c$	0.02		0.10	0.08	0.13	0.40	0.30
	(1.69)		(4.26)	(3.67)	(4.98)	(11.06)	(9.03)
	[0.02]		[0.04]	[0.04]	[0.07]	[0.15]	[0.08]
FF6	0.00	0.08		0.08	0.17	0.41	0.30
	(0.00)	(3.84)		(3.84)	(5.25)	(8.84)	(9.20)
	[0.00]	[0.03]		[0.03]	[0.14]	[0.19]	[0.10]
FF6$_c$	0.02	0.00	0.02		0.08	0.34	0.26
	(1.64)	(0.00)	(1.64)		(3.50)	(8.21)	(8.79)
	[0.02]	[0.00]	[0.02]		[0.08]	[0.17]	[0.10]
Q4	0.06	0.03	0.09	0.06		0.21	0.24
	(2.70)	(2.32)	(3.62)	(3.39)		(6.85)	(7.45)
	[0.05]	[0.03]	[0.07]	[0.04]		[0.10]	[0.06]
BS6	0.20	0.17	0.20	0.17	0.00		0.22
	(5.57)	(5.43)	(5.57)	(5.43)	(0.00)		(8.21)
	[0.13]	[0.10]	[0.13]	[0.10]	[0.00]		[0.07]
DHS	0.13	0.14	0.14	0.15	0.12	0.27	
	(5.11)	(5.63)	(5.23)	(5.75)	(5.03)	(8.28)	
	[0.08]	[0.07]	[0.09]	[0.08]	[0.07]	[0.13]	

Table: 4: **Frontier Expansion** – continued.

Panel B: Net				**Union Model (M1)**			
Base "M0"	FF5	$FF5_c$	FF6	$FF6_c$	Q4	BS6	LSB
FF5		0.03	0.01	0.04	0.05	0.07	0.18
		(1.68)	(0.82)	(1.89)	(2.18)	(3.11)	(4.27)
		[0.03]	[0.02]	[0.04]	[0.05]	[0.06]	[0.06]
$FF5_c$	0.00		0.01	0.01	0.04	0.06	0.15
	(0.02)		(0.71)	(0.70)	(1.67)	(2.62)	(3.98)
	[0.00]		[0.01]	[0.01]	[0.04]	[0.06]	[0.06]
FF6	0.00	0.03		0.03	0.04	0.06	0.17
	(0.00)	(1.62)		(1.62)	(1.87)	(2.71)	(4.16)
	[0.00]	[0.02]		[0.02]	[0.04]	[0.06]	[0.06]
$FF6_c$	0.00	0.00	0.00		0.03	0.05	0.14
	(0.02)	(0.00)	(0.02)		(1.43)	(2.25)	(3.89)
	[0.00]	[0.00]	[0.00]		[0.04]	[0.06]	[0.07]
Q4	0.03	0.05	0.04	0.05		0.04	0.16
	(1.53)	(2.07)	(1.62)	(2.17)		(1.86)	(3.91)
	[0.03]	[0.03]	[0.03]	[0.04]		[0.06]	[0.08]
BS6	0.02	0.03	0.02	0.03	0.00		0.14
	(0.99)	(1.44)	(0.99)	(1.44)	(0.00)		(3.60)
	[0.03]	[0.03]	[0.03]	[0.03]	[0.00]		[0.10]
DHS	0.05	0.06	0.05	0.06	0.05	0.07	
	(1.84)	(2.29)	(1.94)	(2.35)	(2.06)	(2.72)	
	[0.06]	[0.06]	[0.06]	[0.06]	[0.05]	[0.08]	

Table 5: **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: $\Delta\%Sh(M0,M1) = Sh(M0,M1)/Sh(M0) - 1$. Panel A reports the results using gross returns. Panel B 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 till December 2021.

Panel A: Gross	**Union Model (M1)**						
Base Model (M0)	FF5	FF5$_c$	FF6	FF6$_c$	Q4	BS6	LSB
FF5		14.9	15.8	27.4	29.6	84.4	65.1
		(7.8)	(8.6)	(11.3)	(8.7)	(16.6)	(10.5)
FF5$_c$	2.6		13.8	11.4	15.2	64.3	51.6
	(2.9)		(6.1)	(6.0)	(6.9)	(17.3)	(10.8)
FF6	0.0	10.0		10.0	15.9	60.8	45.1
	(0.0)	(4.5)		(4.5)	(11.9)	(24.0)	(10.6)
FF6$_c$	2.1	0.0	2.1		7.3	48.4	38.1
	(2.1)	0.0	(2.1)		(6.6)	(21.5)	(10.2)
Q4	4.7	4.1	7.9	8.0		26.0	33.5
	(3.8)	(3.6)	(5.0)	(6.2)		(12.3)	(7.6)
BS6	19.0	18.2	19.0	18.2	0.0		26.2
	(10.6)	(8.6)	(10.6)	(8.6)	(0.0)		(6.0)
DHS	13.4	16.3	15.1	17.9	13.3	35.4	
	(7.4)	(6.9)	(8.7)	(7.9)	(6.2)	(17.8)	
Panel B: Net	**Union Model (M1)**						
Base Model (M0)	FF5	FF5$_c$	FF6	FF6$_c$	Q4	BS6	LSB
FF5		7.6	2.6	9.5	9.2	16.1	34.6
		(8.2)	(3.9)	(10.1)	(8.8)	(15.8)	(13.8)
FF5$_c$	0.0		1.7	1.7	5.6	11.7	28.3
	(0.1)		(2.7)	(2.7)	(6.8)	(13.6)	(14.9)
FF6	0.0	6.7		6.7	7.0	13.2	32.3
	(0.0)	(7.1)		(7.1)	(7.5)	(15.1)	(15.0)
FF6$_c$	0.0	0.0	0.0		4.5	9.9	27.0
	(0.1)	(0.0)	(0.1)		(6.1)	(13.5)	(15.7)
Q4	5.4	9.7	5.9	10.5		8.1	32.2
	(4.6)	(8.1)	(5.0)	(9.0)		(11.2)	(20.1)
BS6	3.6	7.4	3.6	7.4	0.0		29.2
	(4.2)	(8.5)	(4.2)	(8.5)	(0.0)		(22.6)
DHS	7.5	9.8	8.2	10.4	8.4	14.2	
	(9.1)	(9.3)	(9.2)	(9.5)	(8.9)	(16.0)	

Table 6: **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 2. "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. Panel C presents the in-sample estimates and Panel D 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 till 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 (gross returns)

	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	3.9	0.0	0.4	4.2	0.1	4.4	0.0	0.00	1.34	0.27
FF6	96.1	0.0	35.3	0.0	23.9	2.3	10.7	0.0	0.00	1.47	0.26
$FF5_c$	100.0	64.7	0.0	7.0	35.8	0.8	12.7	0.0	0.08	1.54	0.26
$FF6_c$	99.6	100.0	93.0	0.0	57.2	7.5	23.2	3.5	6.21	1.65	0.25
Q4	95.8	76.1	64.2	42.8	0.0	0.0	18.8	0.0	0.00	1.61	0.34
BS6	99.9	97.7	99.2	92.5	100.0	0.0	65.9	63.6	26.41	2.06	0.52
DHS	95.6	89.3	87.3	76.8	81.2	34.1	0.0	32.9	24.59	1.85	0.28

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

	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	5.5	12.1	4.2	2.9	0.3	3.4	0.0	0.02	1.19	0.29
FF6	94.5	0.0	45.3	14.6	19.0	3.3	8.1	0.4	0.88	1.33	0.27
$FF5_c$	87.9	54.7	0.0	8.6	22.4	1.1	7.1	0.0	0.11	1.36	0.27
$FF6_c$	95.8	85.4	91.4	0.0	42.4	7.4	14.3	1.9	3.49	1.47	0.26
Q4	97.1	81.0	77.6	57.6	0.0	4.0	16.9	0.6	0.86	1.51	0.34
BS6	99.7	96.7	98.9	92.6	96.0	0.0	55.9	54.2	29.43	1.89	0.53
DHS	96.6	91.9	92.9	85.7	83.1	44.1	0.0	42.9	27.87	1.77	0.28

Table 6: **In-sample and Out-of-sample Sharpe ratios** – continued.

	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
Panel C: In-sample estimates (net returns)											
FF5	0.0	7.0	0.0	4.9	44.6	28.8	24.5	0.7	1.16	0.83	0.20
FF6	71.1	0.0	54.2	0.0	59.8	41.9	35.7	3.4	5.02	0.89	0.22
$FF5_c$	55.5	27.5	0.0	7.7	55.0	35.5	31.2	1.0	2.57	0.86	0.21
$FF6_c$	76.8	52.5	70.1	0.0	66.5	48.8	41.5	13.6	18.09	0.92	0.23
Q4	49.6	34.8	39.6	28.4	0.0	0.0	32.0	0.6	1.04	0.86	0.26
BS6	66.1	53.2	58.8	46.0	74.6	0.0	45.9	35.1	40.09	0.97	0.36
DHS	75.4	64.2	68.7	58.4	67.9	54.0	0.0	45.6	37.76	0.95	0.13
Panel D: Out-of-sample estimates (net returns)											
	FF5	FF6	$FF5_c$	$FF6_c$	Q4	BS6	DHS	Best	$\sigma(Best)$	SR	$\sigma(SR)$
FF5	0.0	16.0	23.0	21.9	37.8	31.1	19.3	1.6	1.73	0.70	0.19
FF6	62.2	0.0	58.6	23.4	51.7	43.2	27.9	5.8	7.08	0.75	0.20
$FF5_c$	32.5	23.1	0.0	16.3	39.2	31.4	20.7	0.5	1.81	0.71	0.20
$FF6_c$	59.8	29.1	61.5	0.0	51.9	44.2	28.7	6.0	10.37	0.76	0.21
Q4	56.4	42.9	55.4	43.1	0.0	24.8	28.4	5.5	5.79	0.74	0.25
BS6	63.7	52.0	63.0	50.6	49.7	0.0	38.6	27.5	35.19	0.81	0.34
DHS	80.6	72.0	79.2	71.2	71.5	61.3	0.0	53.1	35.70	0.88	0.12

Table 7: **Liquidity:** This table shows the estimated coefficients obtained from fixed effect regressions where we predict the impact on the liquidity of the long-short portfolio of eight construction choices. The construction choice definitions are the same as in figure 4. Liquidity is calculated following Amihud (2002). Monthly portfolio liquidity levels are constructed using data spanning January 1972 and December 2021. We always include factor- and time fixed effects. 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.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3070	-0.43***							
	(-6.59)							
Financials		0.06						
		(1.04)						
Independent			-0.24***					
			(-3.73)					
Jun_Size				0.07**				
				(2.61)				
Micro					7.32***			
					(17.17)			
Ind_Neutral						-0.11***		
						(-3.42)		
NYSE							-2.93***	
							(-15.86)	
VW								-6.05***
								(-16.79)

Table 8: **Factor exposures:** This table shows the estimated coefficients obtained from fixed effect regressions where we predict the impact on the ex-ante long-short normalized factor exposure of eight construction choices. The construction choice definitions are the same as in figure 4. The normalized factor exposures are calculated for each firm on a monthly basis and aggregated to portfolio level. The normalized firm factor exposure is calculated with: $(Variable - Mean)/(Max - Min)$. Monthly normalized factor exposures are constructed using data spanning January 1972 and December 2021. We always include factor- and time fixed effects. Observations are weighted by factor. Coefficients are multiplied by 100. 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.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3070	-1.26***							
	(-4.83)							
Financials		-0.07						
		(-1.13)						
Independent			0.017					
			(0.62)					
Jun_Size				0.08				
				(1.08)				
Micro					0.58***			
					(4.29)			
Ind_Neutral						-0.42***		
						(-3.29)		
NYSE							-0.85***	
							(-4.62)	
VW								-0.42**
								(-3.04)

Table 9: **Portfolio Sparsity:**

	Panel A: 30-70						**Panel B: 80-20**					
	2nd characteristic			**3rd characteristic**			**2nd characteristic**			**3rd characteristic**		
Factors	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
RMW	0.1	0.0	0.1	0.1	0.0	0.1	24.4	0.1	24.5	23.5	1.0	24.5
RMW_{cp}	0.0	0.0	0.0	0.0	0.0	0.0	9.5	0.0	9.5	8.8	0.6	9.5
CMA	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0
MOM	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1
PEAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FIN	2.9	1.9	4.5	2.9	1.9	4.5	72.3	21.0	82.2	72.4	20.5	82.2
	Panel C: NYSE						**Panel D: NAN**					
	2nd characteristic			**3rd characteristic**			**2nd characteristic**			**3rd characteristic**		
Factors	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
RMW	23.0	0.1	23.1	23.0	0.1	23.1	1.5	0.0	1.5	0.5	0.9	1.5
RMW_{cp}	8.8	0.0	8.8	8.8	0.0	8.8	0.6	0.0	0.6	0.0	0.6	0.6
CMA	0.9	0.0	0.9	0.9	0.0	0.9	0.2	0.0	0.2	0.2	0.0	0.2
MOM	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
PEAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FIN	49.8	14.1	56.3	51.4	12.1	56.3	25.4	8.8	30.4	23.8	10.3	30.4

Table 10: **Portfolio Sparsity:**

	Panel E: Incl. Micro						**Panel F: Ex. Micro**					
	2nd characteristic			**3rd characteristic**			**2nd characteristic**			**3rd characteristic**		
Factors	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
RMW	10.1	0.0	10.1	9.2	0.9	10.1	14.4	0.1	14.5	14.4	0.1	14.5
RMW_{cp}	5.2	0.0	5.2	4.5	0.6	5.2	4.3	0.0	4.3	4.3	0.0	4.3
CMA	0.2	0.0	0.2	0.2	0.0	0.2	0.9	0.0	0.9	0.9	0.0	0.9
MOM	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
PEAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FIN	28.8	7.4	32.7	29.0	6.9	32.7	46.4	15.4	54.0	46.2	15.5	54.0
	Panel G: Independent						**Panel H: Dependent**					
	2nd characteristic			**3rd characteristic**			**2nd characteristic**			**3rd characteristic**		
Factors	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low	High	Low	High-Low
RMW	14.6	0.1	14.6	13.7	1.0	14.6	9.9	0.0	9.9	9.9	0.0	9.9
RMW_{cp}	6.3	0.0	6.3	5.7	0.6	6.3	3.2	0.0	3.2	3.2	0.0	3.2
CMA	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
MOM	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
PEAD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FIN	43.9	14.1	51.0	43.8	13.9	51.0	31.3	8.7	35.8	31.5	8.5	35.8

Table 11: **Firms in Q factor portfolios:** This table shows the average number of positions for Q-factor portfolios, averaged across all 256 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 spanning January 1972 and December 2021.

Choice	BHH	BHN	BHL	SHH	SHN	SHL	BLH	BLN	BLL	SLH	SLN	SLL
Excl Micro	81	101	49	81	120	83	50	90	61	48	106	109
Incl Micro	110	134	62	137	210	170	58	105	73	85	190	272
Dep	84	115	66	124	168	112	55	106	86	68	142	158
Ind	106	121	46	94	162	140	54	89	48	65	153	222
20-80	61	116	38	72	166	92	34	93	44	45	148	142
30-70	130	119	74	146	164	161	74	102	90	88	148	238
NAN	125	154	74	92	140	102	74	134	91	64	152	170
NYSE	66	81	38	126	190	151	34	61	43	69	143	210

Figure 7: **Gross Sharpe ratio variation for the Q-factor model:** This figure plots the distribution of annualized gross Sharpe ratios for the Q-factor returns, 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 2. "233" ("23") denotes that the factors are constructed using a 2x3x3 (2x3) sorting methodology. Monthly Q-factor returns are constructed using data spanning January 1972 and December 2021.

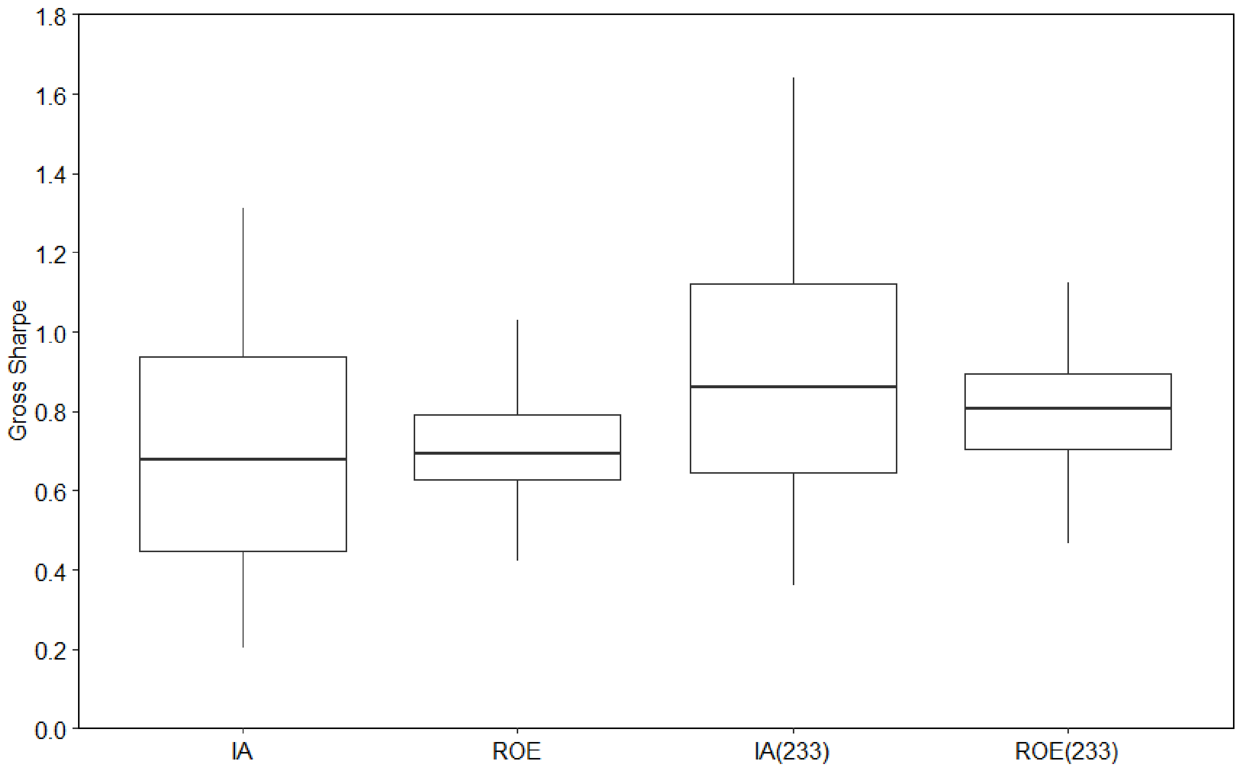
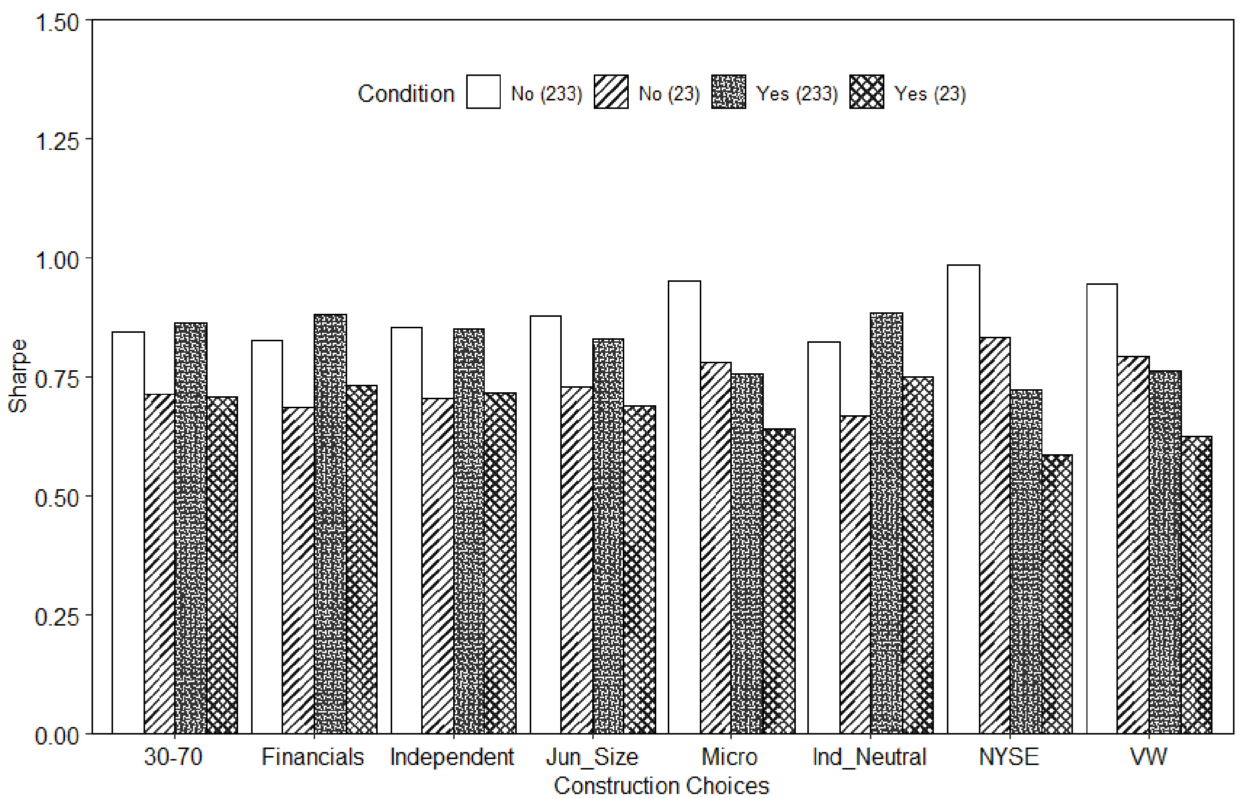


Figure 8: **Construction choices and gross Sharpe ratios:** This figure shows the impact of construction choices on the Sharpe ratio averaged over factors. Sharpe ratios are computed on a gross-basis and are annualized. We consider eight choices. "30-70" refers to the use of the 30th and 70th percentile as breakpoints in the sorting procedure ("Yes") or the use of the 20th and 80th percentile ("No"). "Financial" means that companies in the finance sector are included ("Yes") or excluded ("No") in the factor construction. "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"). "Micro" indicates whether we exclude stocks with the smallest 20% market cap based ("Yes") or not ("No"). "Neutral" means that portfolio sorts are constructed using industry-adjusted characteristics ("Yes") or the standard characteristics ("No"). NYSE indicates whether the NYSE cross-section is used to construct breakpoints ("Yes") or the full NYSE-AMEX-Nasdaq cross-section ("No"). "VW" indicates whether factors are calculated by using value-weighting ("Yes") or equal-weighting ("No"). "233" ("23") denotes that the factors are constructed using a 2x3x3 (2x3) sorting methodology. Monthly Q-factor returns are constructed using data spanning January 1972 and December 2021.



8 Appendix: for online publication only

A.1: Sorting variables

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 $t2$ to the fiscal year ending in $t1$ divided by total assets at $t2$. 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 $t1$ 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):

$$CAR_i = \sum_{d=-2}^{d+1} (R_{i,d} - Rm, d) \tag{16}$$

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.

A.2: Additional results

Figure 9: **Gross Sharpe ratio variation within equal-weighted factors:** This figure plots the distribution of annualized gross Sharpe ratios for equal-weighted long-short factor returns, 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 2. The sample, to calculate these Sharpe ratios, runs from January 1972 till December 2021.

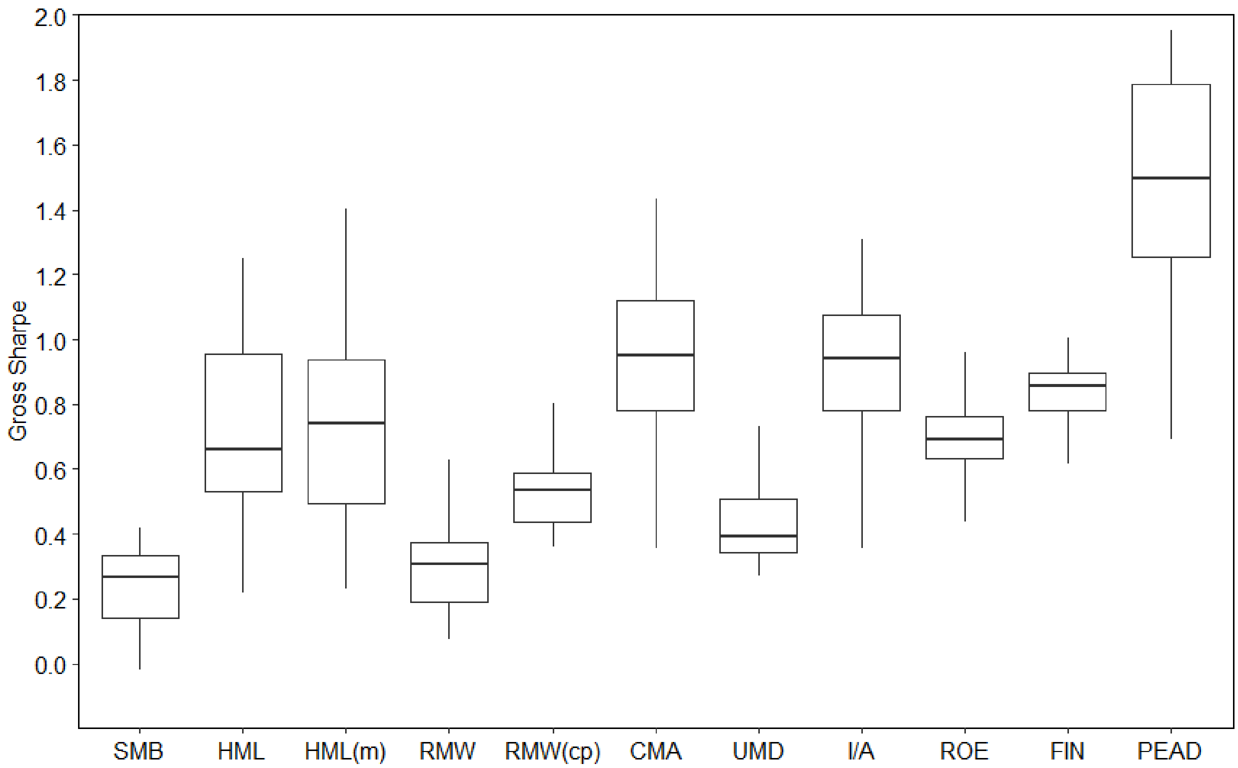


Figure 10: **Net Sharpe ratio variation within equal-weighted factors:** This figure plots the distribution of annualized gross Sharpe ratios for equal-weighted long-short factor returns, where a factor is constructed N times by using the N different factor construction methods. The black solid line 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 2. The sample, to calculate these Sharpe ratios, runs from January 1972 till December 2021.