University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 463

Low Risk, High Variability: Practical Guide for Portfolio Construction

Antonello Cirulli, Gianluca De Nard, Joshua Traut and Patrick Walker

Revised version, November 2025

Low Risk, High Variability: Practical Guide for Portfolio Construction

Antonello Cirulli [1] Gianluca De Nard [2] Joshua Traut [3] Patrick Walker [4]

November 1, 2025

Abstract

This paper explores how various portfolio construction choices influence the performance of low-risk portfolios. We show that methodological decisions critically influence portfolio outcomes, causing substantial dispersion in performance metrics across weighting schemes and risk estimators. This can only be marginally mitigated by incorporating constraints such as short-sale restrictions and size or price filters. Our analysis reveals that volatility-based estimators yield the most favorable performance distribution, outperforming beta-based approaches. Transaction costs are found to significantly affect performance and are vitally important in identifying the most attractive portfolios, highlighting the importance of realistic implementation constraints. Through rigorous empirical analysis, this study bridges the gap between theoretical insights and practical applications, offering actionable guidance to investors on limiting nonstandard errors.

Keywords: Low-Risk Investing, Methodology, Nonstandard Errors, Portfolio Construction.
JEL Classification: C52, G11, G12, G15, G17.

[1]OLZ AG, Zurich, Switzerland | antonello.cirulli@olz.ch
[2]Department of Economics, University of Zurich, Switzerland; University of Liechtenstein, Liechtenstein & OLZ AG, Zurich, Switzerland | gianluca.denard@econ.uzh.ch | https://denard.ch
[3]Swiss Institute of Banking and Finance, University of St. Gallen, Switzerland & BIT Capital GmbH, Berlin, Germany | joshua.traut@unisg.ch | https://www.joshuatraut.com
[4]Department of Finance, University of Zurich, Switzerland & OLZ AG, Zurich, Switzerland | patrick.walker@olz.ch

1 Introduction

Methodological uncertainty has emerged as a critical topic in quantitative finance, particularly in the study of portfolio construction and performance evaluation. The growing complexity of financial datasets, the large number of reported factors that appear relevant (Cochrane, 2011), the p-hacking problem (Harvey, 2017), and the proliferation of portfolio design choices have led to concerns about the robustness and replicability of research findings (e.g., see Hou et al., 2020; Jensen et al., 2023). These issues are especially relevant in the context of factor investing, where methodological choices can significantly affect the performance of strategies that seek to capture anomalies such as value, quality, momentum, and low-risk premia (Hasler, 2023; Walter et al., 2023; Soebhag et al., 2024).

The existing literature highlights the challenge posed by methodological uncertainty in financial research. Menkveld et al. (2024) propose the concept of "nonstandard errors", a measure of variability introduced by various methodological decisions. They show that these errors could be as large as or even larger than traditional standard errors. Soebhag et al. (2024) apply the notion of nonstandard errors to factor sorting portfolios, suggesting that conclusions drawn from portfolio backtests may often depend more on design choices than on underlying economic phenomena. Similarly, Hasler (2023) and Walter et al. (2023) emphasize the sensitivity of factor portfolio results to seemingly minor methodological details, while Chen et al. (2024) illustrate the impact of decision variables in machine learning stock prediction models. In this portfolio context, methodological uncertainty arises from a variety of choices, including the selection of exclusion criteria (e.g., size, price, sector and characteristics filter), the rebalancing frequency, portfolio size or weighting scheme. These decisions often involve trade-offs between theoretical rigor and practical implementability.

Although the current literature emphasizes methodological uncertainty in the portfolio construction process, we find that not enough effort has been put in quantifying the effect of decision variables related to the definition of the sorting variable – arguably the most relevant input. We address this gap by examining the computation of a specific factor theme, and thus quantify the estimation error of the signal used to construct sorting portfolios. Contrary to existing literature that takes factor definitions provided by the corresponding original papers, we show that choices in factor computation – such as the specific estimator, parametrization, and lookback window length – are equally critical and should not be overlooked. For instance, replicating a low-volatility portfolio involves decisions about the type of volatility estimator to use, including variations in lookback periods and data weighting. In this study, we demonstrate that the impact of these (low-risk) factor computation choices is substantial and comparable to the influence of portfolio sorting construction decisions highlighted by Hasler (2023), Walter et al. (2023), and Soebhag et al. (2024).

To demonstrate that the above mentioned extensions lead to substantial variation in factor portfolio performance, we illustrate the impact of construction choices on the cumulative (net of two-way transaction costs) returns of low-risk portfolios in Figure 1. Each path corresponds to a specific long-only portfolio construction method, as detailed below. In orange we plot 1,620 equally-weighted portfolios, in blue 1,620 value-weighted portfolios, and in black a value-weighted market portfolio

that serves as our benchmark. Despite our focus on one factor theme, the figure reveals significant dispersion in portfolio performance, underscoring the critical influence of methodological decisions and the necessity for a deeper evaluation of decision criteria. The compound annual growth rate of low-risk portfolios ranges from -3.9% to +14.0%, compared to +11.9% of the benchmark.

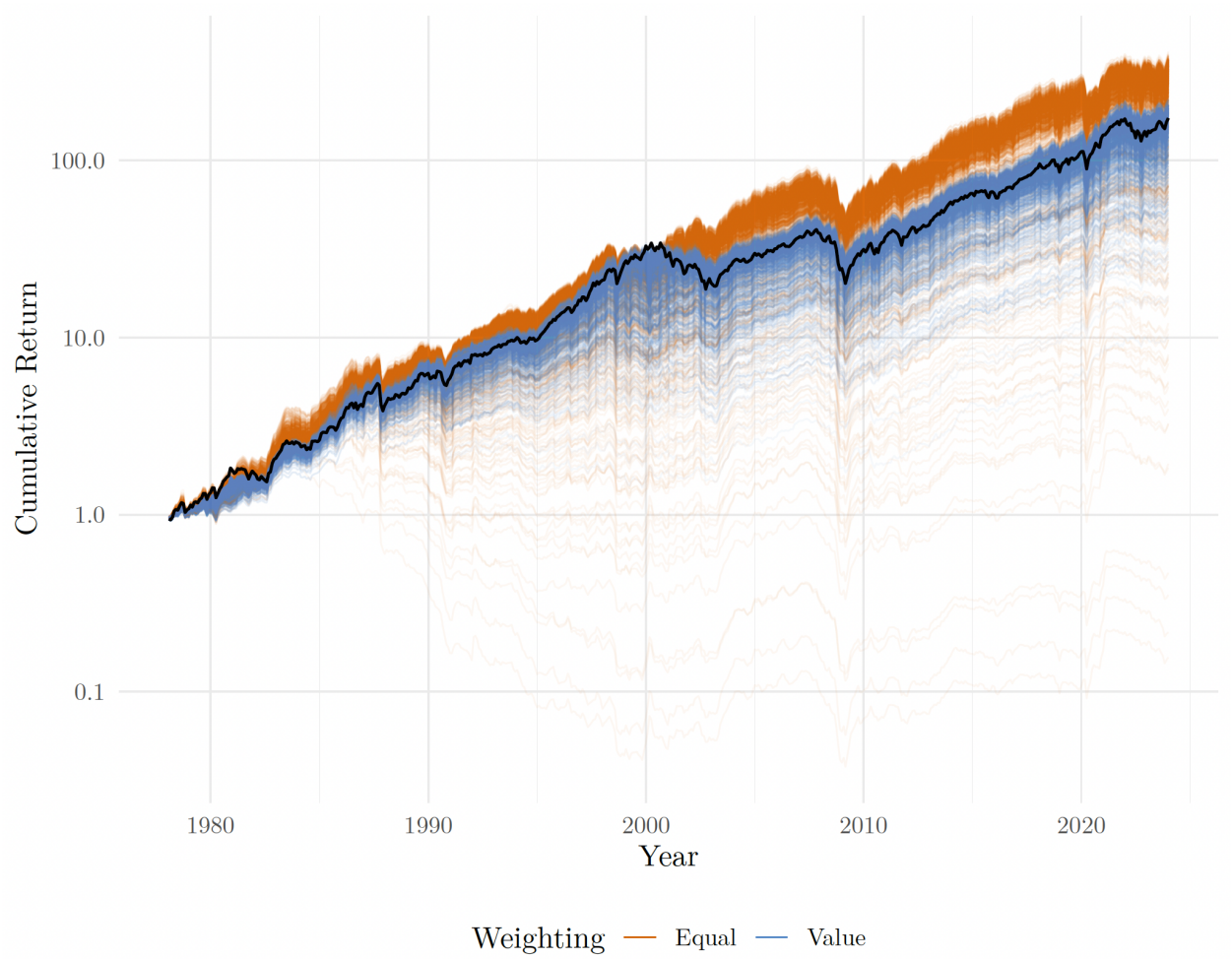

Figure 1: Cumulative returns of long-only low-risk portfolios with transaction costs

The figure shows the dispersion of cumulative returns of long-only portfolios after transaction costs across portfolio construction decisions. In total, the plot contains returns for 3,240 portfolios. We distinguish between portfolios with equally-weighted stocks (orange) and value-weighted stocks (blue). The cumulative returns are calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023 and presented on a log-10 y-axis.

We extend methodological uncertainty to the sorting variable and broaden the evidence by emphasizing practically relevant portfolio designs. Whereas much of the literature evaluates dollar-neutral long–short portfolios and ignores transaction costs, we examine more realistic, implementable strategies – specifically long-only and fully invested, limited-short portfolios. We also analyze turnover and report performance both before (gross) and after (net) accounting for transaction costs. To fully assess the dispersion of the results, we evaluate performance across multiple relevant metrics rather than relying on a single measure and also compare them with a benchmark,

which is neglected in the current literature. This approach provides a more holistic view of the influence of portfolio construction choices.

To investigate these extensions on methodological uncertainty and nonstandard errors in portfolio construction, we focus on the "factor zoo" within a single factor theme rather than across factors. Specifically, we focus our empirical analysis on low-risk portfolios to examine the dispersion within this factor. For many fundamental factors, such as value or quality, certain proposed methodological enhancements, such as varying the length of the lookback window or employing alternative estimation techniques, are often not applicable or meaningful due to the low time series granularity of the data, stemming from infrequent updates in accounting cycles. This renders all accounting-based factors unsuitable for our purposes.

The need for a comprehensive study of low-risk strategies is underscored by the growing attention they have received over the past two decades, both in academia and industry. This interest is driven by the apparent paradox that less volatile stocks often deliver higher risk-adjusted returns, a phenomenon widely referred to as the low-risk anomaly (see Blitz et al., 2023; Traut, 2023; Soebhag et al., 2025, for an overview including a wide range of volatility premiums). Recent investment manager surveys highlight that low-risk investing is among the most commonly employed price-based anomaly strategies, on par or even surpassing momentum (Invesco, 2023, 2024). While extensive academic literature supports the existence of this anomaly across global (equity) markets (Blitz & van Vliet, 2007; Blitz et al., 2020), there remains a gap between theoretical insights and their practical implementation.

This paper bridges that gap by focusing on the impact of construction and evaluation choices of realistic and implementable low-risk portfolios for U.S. equities. Most existing studies focus on simplistic sorting mechanisms and overlook the profound influence of methodological choices, such as risk estimator selection and configuration, transaction costs handling, rebalancing frequency, and portfolio constraints. These decisions are pivotal in shaping portfolio performance, yet their impact remains largely underexplored both in the context of low-risk investing and factor investing more broadly.

Our results indicate that nonstandard errors are more modest for long-only investors, suggesting that low-risk investing is relatively robust in practice. The most consequential modeling decision is the portfolio weighting scheme: equal-weighting versus value-weighting has the largest impact on outcomes. By comparison, the choice of risk estimator is secondary but still material. Together, these findings provide encouraging evidence for the reliability of long-only low-risk strategies while clarifying where specification choices matter most.

2 Data and Methodology

2.1 Data

We construct our sample using the CRSP daily dataset, spanning from January 1973 to December 2023. The dataset is restricted to primary U.S. stocks listed on AMEX, NYSE, and NASDAQ, excluding stock-days with missing observations for returns or shares outstanding. Since the longest estimation window in our analysis is five years, we start our out-of-sample analysis in January 1978, providing 11,598 daily returns per portfolio and benchmark. This starting point is selected to avoid the structural changes introduced in the early 1970s, as the CRSP universe tripled in size in 1973. The eligible universe for out-of-sample analysis ranges in size from 3,564 to 7,465 stocks over time with an average of 5,089 stocks.

Daily data for the Fama and French (2015) five-factor model are sourced from the WRDS database, while NBER recession identifiers are obtained from the Federal Reserve Bank of St. Louis.

Transaction costs are approximated using half spreads, as proposed by Novy-Marx and Velikov (2016). When quote data are available, the spreads are calculated as the quoted bid-ask spread divided by twice the contemporaneous mid-point, averaged over month m. In cases where quote data are unavailable, spreads are estimated using the CHL spread estimator by Abdi and Ranaldo (2017). For any remaining missing values, we impute spreads using the methodology described in Novy-Marx and Velikov (2016).

2.2 Portfolio Construction and Evaluation Choices

In our empirical analysis we focus on some of the most prominent estimators in the low-risk literature. First, we use the total volatility of historical returns (*hist*) as advocated by Blitz and van Vliet (2007), Baker and Haugen (2012), and Blitz et al. (2013). This measure is arguably the simplest in our analysis as it is based on the sample volatility, i.e. the standard deviation of simple returns, $\hat{\sigma}_{i,T}^{hist}$, for all stocks in the eligible universe $i = 1, \ldots, N$.

Secondly, we employ the seminal RiskMetrics method to estimate the risk of a stock (RiskMetrics, 1996; Mina & Xiao, 2001; Zumbach, 2007). It is closely related to the *hist* methodology but places more weight on more recent observations in the calculation of volatility. Henceforth, $\sigma_{i,T}^{ewma}$ is estimated using an exponential weighted moving average (*ewma*) as

$$\hat{\sigma}_{i,T}^{ewma} = \sqrt{\lambda(\hat{\sigma}_{i,T-1}^{ewma})^2 + (1-\lambda) r_{i,T-1}^2}\,, \tag{1}$$

where $\lambda \in [0, 1]$ is the decay factor.

Another refinement of the *hist* methodology is to use idiosyncratic volatility of residual returns instead of total returns to determine risk. This is commonly done relative to the Fama and French (1993) three-factor model as proposed by Ang et al. (2006), which was later validated by Ang et al.

(2009) and Dimson et al. (2017), and can be computed by performing the following regression

$$r^e_{i,t} = \hat{\alpha}_i + \hat{\beta}_{i,m} r^e_{m,t} + \hat{\beta}_{i,SMB} r_{SMB,t} + \hat{\beta}_{i,HML} r_{HML,t} + \hat{\epsilon}^{ff3}_{i,t} \ , \text{ for } t = 1, \ldots, T, \tag{2}$$

where m, SMB, and HML refer to the three factors of the Fama and French (1993) three-factor model and the superscript e denotes returns in excess of the risk-free rate, r_f. The standard deviation of $\hat{\epsilon}^{ff3}_{i,t}$, denoted as $\hat{\sigma}^{ff3}_{i,T}$, is then taken as the risk measure. Thus, we refer to this risk estimator as *ff3*. We also include the idiosyncratic volatility of residuals in the Fama and French (2015) five-factor model (*ff5*) as it represents a more contemporaneous methodology that is likely employed by practitioners to isolate idiosyncratic risk.

Another prevalent research stream focuses on measuring risk with market beta (Fama & MacBeth, 1973; Asness et al., 2014; Frazzini & Pedersen, 2014). We follow this stream and calculate our *beta* risk measure as

$$\hat{\beta}_{i,m} = \hat{\rho}^e_{i,m} \frac{\hat{\sigma}^e_i}{\hat{\sigma}^e_m} \ , \tag{3}$$

where $\hat{\rho}^e_{i,m}$ is the sample correlation of r^e_i and r^e_m that is calculated from overlapping three-day log returns, $r^{3d}_{i,t} = \sum_{k=0}^{2} \ln(1 + r_{i,t+k})$, to account for nonsynchronous trading. $\hat{\sigma}^e_i$ and $\hat{\sigma}^e_m$ are the sample standard deviations of r^e_i and r^e_m respectively that are calculated from daily log returns directly.

For our risk estimation described above and for portfolio construction, we require a minimum of 90% of daily return observations for the respective lookback period being available, and we filter out stocks with missing return on the last day of our lookback period. Furthermore, we exclude stocks that have missing volume observations for the last month in the portfolio formation. [1]

Regarding our choices with respect to the methodological setup of portfolio construction, we are strongly guided by Hasler (2023), Walter et al. (2023), and Soebhag et al. (2024). We exclude decision forks that are commonly used in studies on factors that are based on book variables (e.g., the exclusion of financial or utility companies) as the rationale for these decisions does not apply to low-risk investing. Furthermore, we expand their decision variables by choices that may not be used in academic research yet but are relevant for practitioners.

1. **Portfolio Type:** Current studies that examine the influence of methodological choices in the portfolio construction context focus solely on studying long-short factor portfolios. These portfolios are not only difficult (or even impossible) to implement for many market participants but also involve taking on leverage, which could be seen as contradictory to the goal of investing with low risk. However, the recent work by Blitz et al. (2024) shows that taking

[1] The missing data filter ensures consistency in the lookback period; for instance, a stock with only four years of data is excluded from the 5-year lookback portfolio. While the literature is generally less strict (e.g., Hou et al., 2020), we impose higher data coverage standards to ensure practical relevance. The volume filter serves a similar purpose, mitigating known data issues in the early CRSP daily files (Bryzgalova et al., 2024) and preventing bias from erroneous records.

on some leverage can significantly improve the performance of low-risk portfolios. Because of this, we consider three portfolio types in our analysis: a long-only portfolio, which we denote as *100-0*; a fully-invested limited-short portfolio, that takes on a leveraged position of 130% on the least risky stocks and finances this by short selling 30% of the riskiest stocks (*130-30*); and the standard dollar-neutral long-short portfolio that goes long on the least risky stocks and short on their most risky counterparts with equal investment amounts (*100-100*).

2. **Size Exclusion:** To mitigate the effects of small and potentially illiquid stocks on the results, we employ the two most commonly used size filters as in Walter et al. (2023) and Soebhag et al. (2024). These filters either exclude stocks that are smaller than either the 10%, or 20% thresholds based on the market capitalization of NYSE stocks.

3. **Price Exclusion:** Price exclusions aim to reduce the impact of trading frictions like short sale constraints on the outcome of the portfolios. Again, we focus on the most popular choices as identified by Hasler (2023), Walter et al. (2023), and Soebhag et al. (2024) and exclude stocks with a price below either 1$, or 5$ in the portfolio formation.

4. **Lookback Window:** The lookback for the case of low-risk portfolios specifies how much past data (T) are used to estimate the respective risk measure for the portfolio formation. This is an important choice as one needs to balance the trade-off between more timely risk estimates and having too few data points. The low-risk literature shows considerable variation in the choice of lookback windows, spanning from as short as one month (e.g., Ang et al., 2006) to as long as five years (e.g., Frazzini & Pedersen, 2014). To keep the analysis manageable, we limit our focus to three specific settings, using data from the past one, three, and five years to compute our risk estimators.[2]

5. **Rebalancing Frequency:** We include monthly and annual rebalancing frequencies, as these are the two most common choices in academic literature. To capture further relevant frequencies, we consider weekly and quarterly rebalancing.

6. **Portfolio Buckets:** We include groups of three (tercile), five (quintile), and ten (decile) portfolios.

7. **Weighting Scheme:** The weighting scheme is among the most frequently controlled decision criteria. Therefore, same as in Hasler (2023), Walter et al. (2023), and Soebhag et al. (2024), we consider both equal and value weighting.

Unlike current studies that primarily address methodological uncertainty in portfolio construction, we expand our analysis to include decisions related to portfolio evaluation. Although existing studies often focus solely on average returns (Hasler, 2023; Walter et al., 2023) or Sharpe ratios (Soebhag et al., 2024), we evaluate our portfolios using a range of criteria (based on daily frequency):

[2]It is important to note that for the *ewma* estimator, the decay factor, λ effectively sets the lookback window for the analysis. Because of this, we choose λ for the respective lookback windows such that 99.9% of the probability mass falls within one, three and five years, leading to λ of 0.97, 0.99 and 0.995 respectively.

Cumulative return (Cum.), average return (Ret.), alpha, generalized alpha[3], standard deviation (Std.), Sharpe ratio (SR), information ratio (IR) and maximum drawdown (m. DD).[4]

We calculate all performance measures for the total period of our sample, that is, we include all the 11,598 daily (stock, benchmark, and portfolio) returns from January 1978 to December 2023. In addition, we compare them with a benchmark based on the value-weighted market return of the Fama and French (2015) five-factor model.

Last, building on findings by Detzel et al. (2023), who demonstrate that the performance of factor models is significantly affected by the inclusion or exclusion of transaction costs, we consider two cases for each portfolio configuration. The first excludes transaction costs, as is common in low-risk investing literature, while the second includes them to evaluate whether low-risk strategies hold up under realistic implementation assumptions.

Taking into account all feasible portfolio constructions and performance evaluation combinations, we examine a total of 136,080 performance metrics for 9,720 portfolios. Our complete empirical setup is illustrated as a decision tree in Figure 2.

To ground the analysis in practice, we show how arguably the most prominent U.S. low-volatility index aligns with these choices: The S&P 500 Low Volatility Index measures the performance of the 100 least volatile stocks in the S&P 500. The index benchmarks low volatility strategies for the U.S. stock market. Along the construction choices of Figure 2 it is constructed based on the historical volatility risk estimator, it is a long-only portfolio, it is based on the S&P 500 universe, the lookback window is one year, it has quarterly rebalancing frequency, it computes five buckets and invests in the 100 least volatile stocks, it weights them relative to the inverse of their corresponding volatility, with the least volatile stocks receiving the highest weights. Note that this inverse-volatility weighting is not the standard in the academic factor literature, where equal- or value-weighted portfolios are more common, but it is a practical choice for index construction and implementation.

3 Empirical Results

3.1 General Overview of the Portfolio Strategies

To obtain a general understanding of the impact of construction choices on low-risk (sorting) portfolios, we first focus on the dispersion and similarity of all the investigated strategies. We begin by examining whether Sharpe ratios are robust to varying construction choices. Note that the results for other performance measures, such as cumulative returns, mean returns, standard deviation, and maximum drawdown, are comparable and are available upon request. Therefore,

[3]We employ the generalized alpha measure by Novy-Marx and Velikov (2016) as it offers a more reliable evaluation of risk-adjusted returns in the presence of transaction costs than standard regression alphas.

[4]It is worth noting that, while calculating cumulative returns for long-only and limited-short portfolios is straightforward, there is no consensus on how they should be calculated for dollar-neutral long-short portfolios. Following Detzel et al. (2023), we adopt an interest-free cash account methodology.

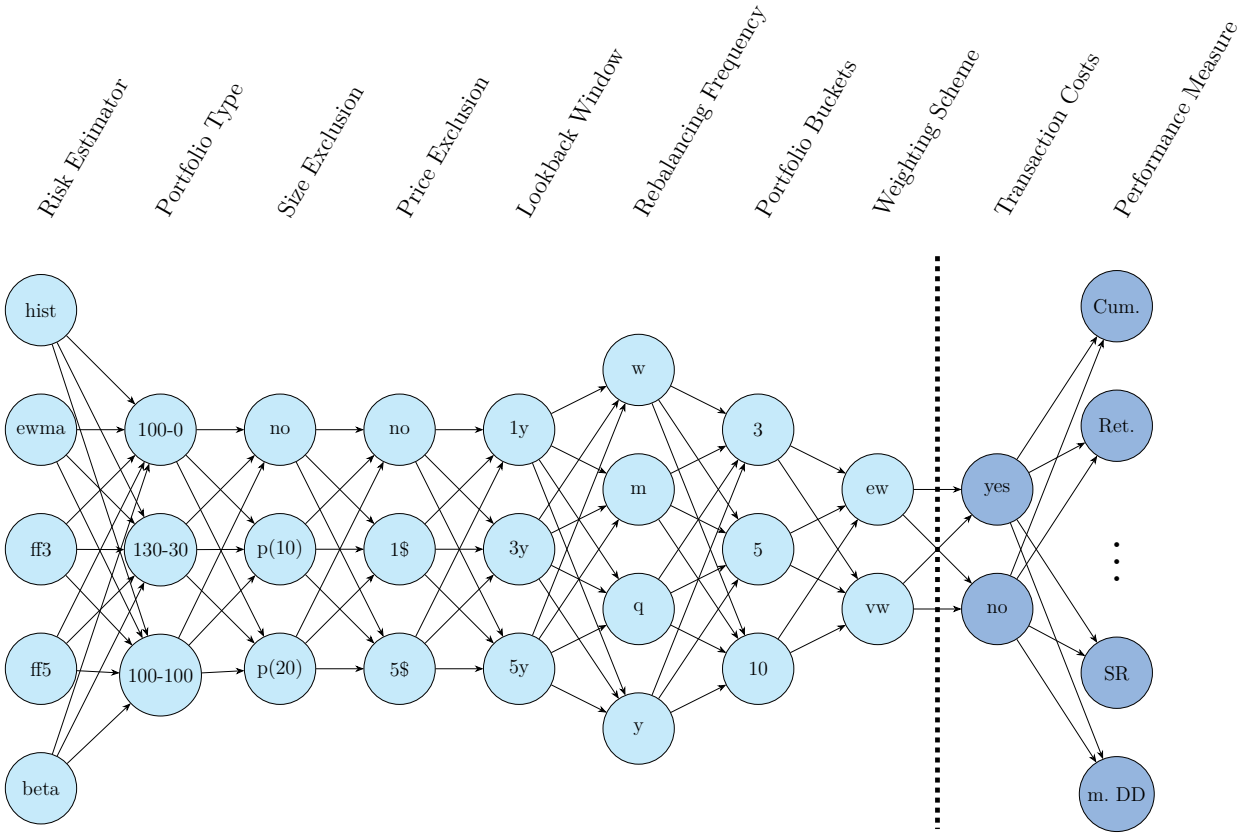

Figure 2: Decision tree of portfolio construction and evaluation choices

The figure illustrates all possible decision paths for our portfolio construction and evaluation. We consider the paths of eight portfolio construction decision forks and two performance evaluation forks for our sorting portfolios. The portfolio construction forks include the estimation of risk (historical volatility, exponentially weighted moving average, idiosyncratic volatility with respect to the Fama and French 3-factor and 5-factor model, or market beta), portfolio type (long-only (*100-0*), limited-short (*130-30*), or long-short (*100-100*)), small stocks exclusion dependent on market equity quantiles (none, smaller than p(10) or p(20)), stock price exclusion (none, smaller than \$1 or \$5), lookback window for the risk estimators (past returns of the previous one, three, or five years), the number of sorting portfolio buckets (three, five, and ten), and weighting scheme (equal- or value-weighting). The performance evaluation forks include transaction costs (with or without) and the performance metric (cumulative return, average return, alpha, standard deviation, Sharpe ratio, information ratio, or maximum drawdown). In total we construct 9,720 portfolios for which we calculate 136,080 performance metrics.

we restrict our attention to risk-adjusted performance measured by the Sharpe ratio. However, it should be noted that whenever returns are negative the Sharpe ratio cannot not be used for ranking. Nevertheless, the dispersion of Sharpe ratios is still informative about how portfolio construction decisions affect risk-adjusted performance.

Since many investors face short-sale restrictions and other regulatory limits, long-only implementations of equity factors are the default choice for a broad segment of the investor base (Baker et al., 2011; Frazzini & Pedersen, 2014). According to various estimates, at least 90% of global stocks are invested in a long-only way, underscoring the importance of evaluating realistically implementable

portfolios with no (or limited) short-selling. For this reason, the main text focuses on portfolios that are arguably the most relevant from a practical investor perspective – namely, long-only (*100-0*) and limited-short (*130-30*). For completeness, we also report results for dollar-neutral long-short portfolios (*100-100*) in Appendix A.

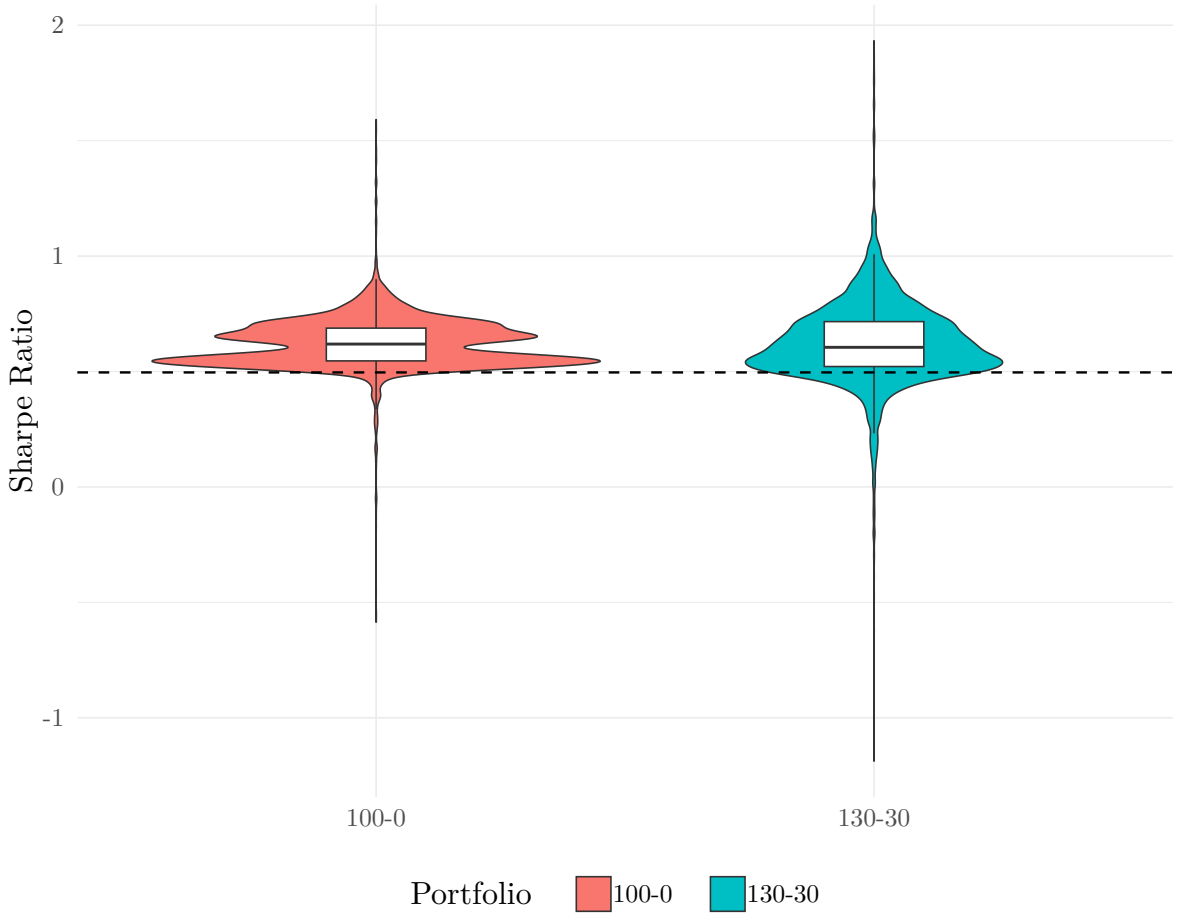

Figure 3: Sharpe ratios of portfolios across portfolio construction choices

This figure plots the distribution of (annualized) gross (without transaction costs) and net (with transaction costs) Sharpe ratios for all the 6,480 portfolios, respectively $3,240 \times 2$ Sharpe ratios of long-only (*100-0*) and $3,240 \times 2$ of limited-short (*130-30*) portfolios across varying portfolio construction choices. The Sharpe ratio of our benchmark value-weighted market portfolio is shown as a dashed line. The boxes represent the interquartile range and the median of the distribution. All the portfolios are based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

In Figure 3, we plot the annualized Sharpe ratios with and without transaction costs for the long-only portfolios (*100-0*) and of limited-short portfolios (*130-30*). The benchmark is the Sharpe ratio of the market-cap-weighted portfolio, indicated via a dashed line.

Figure 3 reveals substantial variation in the annualized Sharpe ratios obtained. Depending on how the low-risk portfolio is constructed and evaluated, the Sharpe ratios range between -0.59 and 1.59 for *100-0*, and between -1.19 and 1.93 for *130-30*. While both portfolios exhibit similar average

and median Sharpe ratios, slightly above 0.6, the figure indicates that the uncertainty arising from construction choices is larger for portfolios that permit short selling. Another notable finding is the difference in relative performance against the benchmark in terms of Sharpe ratio. While 97% of the *100-0* portfolios outperform the market-cap benchmark Sharpe ratio of 0.5, "only" 83% of the *130-30* portfolios outperform the benchmark. Interestingly, we find that the dispersion within dollar-neutral long-short portfolios is even larger and that only 17% of *100-100* portfolios achieve a positive Sharpe ratio; see Figure A.1.

Another perspective on the portfolio construction uncertainty problem is to evaluate the similarity in performance across strategies. In Figure 4, we report the average correlation between two portfolios, of which one is based on an estimator belonging to the class on the corresponding row, while the other one is based on an estimator from the class on the corresponding column.[5] On the diagonal we present the average correlation between two portfolios based on estimators within the same estimator class. The reported correlation numbers are for portfolios that include transaction costs; however, the gross results are nearly identical. Since we consider different specifications of low-risk portfolios, we find that correlations are generally high. Nevertheless, several noteworthy patterns emerge across portfolio types and risk estimators.

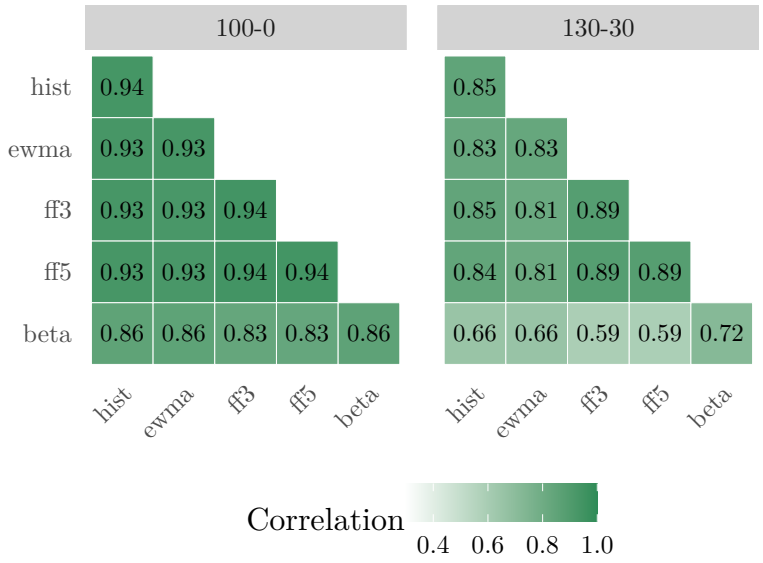

Figure 4: Average correlations between and within an estimator class including transaction costs

This figure shows the average correlation between two portfolios, of which one is based on an estimator belonging to the class on the corresponding row, while the other one is based on an estimator from the class on the corresponding column. In particular, the diagonal elements are average correlations between two portfolios based on estimators within a certain class. The analysis is carried out separately for long-only (*100-0*) and limited-short (*130-30*) portfolios. Correlations are calculated based on 11,598 daily out-of-sample returns including transaction costs from January 1978 to December 2023.

First, consistent with the findings on Sharpe ratio dispersion, we observe that correlation num-

[5]Correlations for long-short portfolios are reported in Figure A.2 in the Appendix.

bers decrease if short selling is permitted. For the *100-0* portfolios, the average correlation is 0.91 and for *130-30* it drops to 0.78. Second, the correlations between portfolios for a given risk estimator (diagonals) are generally higher than the correlations between different risk estimators (off-diagonals). Third, portfolios sorted on *beta* exhibit the lowest correlations, both with portfolios based on other risk estimators and with other *beta* portfolios. This highlights the instability of *beta* as a portfolio construction variable, given that it relies on an OLS regression coefficient, which is susceptible to potentially large estimation errors. However, *beta*-based portfolios also offer the largest diversification potential when combined with other low-risk portfolios.

Analyzing the out-of-sample distribution of various portfolio performance measures and the correlations between portfolio returns, we conclude that increasing the degrees of freedom in portfolio construction exacerbates the problem of methodological uncertainty. This further supports the idea that weight constraints in portfolio construction are beneficial to the robustness of a strategy and can help mitigate the dispersion of the results.[6] To quantify the methodological uncertainty of various practically relevant low-risk portfolios, we focus on the concept of nonstandard errors in the next section.

3.2 Nonstandard Errors Versus Standard Errors

To evaluate the influence of construction choices on portfolio performance and to make our results comparable with existing literature, we calculate standard and nonstandard errors for our collection of low-risk portfolios. In our definition of nonstandard errors, we are guided by Soebhag et al. (2024) who use Sharpe ratios as their evaluation criterion. We define nonstandard errors as the cross-sectional standard deviation of Sharpe ratios across all portfolios (1,296 in total) for a given low-risk estimator and portfolio type. Standard errors, in contrast, are calculated as the average monthly Sharpe ratio (time-series) standard deviation across all construction choices for the same estimator and portfolio type.[7] The annualized results of both error types are shown in Figure 5.

Similar to the findings of Menkveld et al. (2024) on nonstandard errors in research design choices and those of Soebhag et al. (2024) and Walter et al. (2023) on nonstandard errors in factor portfolio sorts, we find that nonstandard errors for low-risk portfolios are generally too large to be ignored and are substantial relative to standard errors. A key contribution of our analysis is the distinction between portfolio types. While the existing literature predominantly focuses on long-short dollar-neutral portfolios, we demonstrate that for long-only portfolios and fully-invested portfolios with limited short selling, nonstandard errors tend to be smaller and are often lower than their standard error counterparts.[8] Overall, our findings indicate that nonstandard errors are less stable across portfolio types and risk estimators compared to standard errors. Furthermore, as the degrees of freedom increase (i.e., more short selling allowed), the relative impact of nonstandard errors (compared to

[6] See, for example, Jagannathan and Ma (2003), who demonstrate that imposing (even incorrect) weight constraints reduces risk in large portfolios and prove that it is equivalent to using a shrinkage estimator for the covariance matrix, as in Ledoit and Wolf (2004) and De Nard (2022).

[7] For the average standard error definition see Menkveld et al. (2024) and Walter et al. (2023).

[8] We report standard and nonstandard errors for long-short portfolios in Figure A.3 in the Appendix.

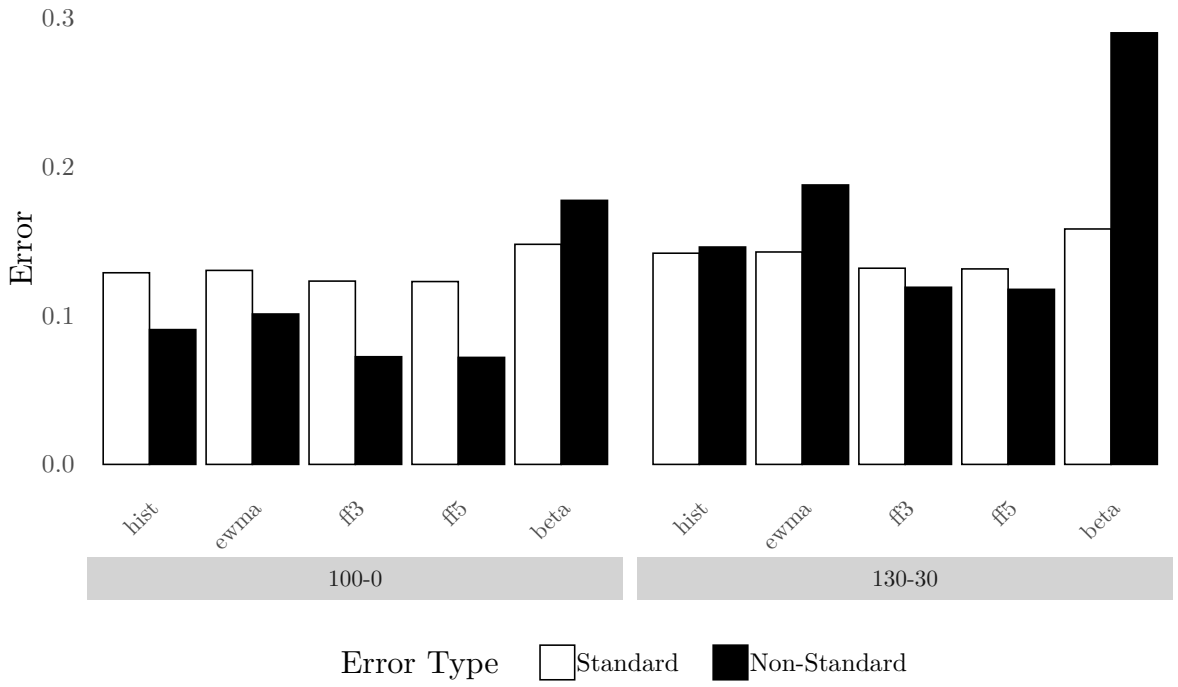

Figure 5: Standard vs. nonstandard errors

This figure plots the nonstandard error (black) and standard error (white) for each risk estimator and portfolio type. In particular, for a given risk estimator and portfolio type, we define the nonstandard error as the standard deviation of the Sharpe ratios across all portfolios of that type, each calculated using the specified risk estimator. In contrast, the standard error is defined by first computing the time series of monthly Sharpe ratios for each portfolio and then calculating its standard deviation. The standard error for a risk estimator and portfolio type is the average of these standard deviations across all portfolios of that type, each based on the given risk estimator. The results are grouped by portfolio type: long-only (*100-0*) and limited-short (*130-30*). Both error types are annualized and calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

standard errors) becomes larger. This finding offers promising implications for investors who face various (regulatory or risk management) constraints on their portfolio holdings. For these investors, nonstandard errors pose less of a challenge than suggested by previous literature.

To evaluate how much each construction choice of our decision tree influences portfolio outcomes we follow Walter et al. (2023) and calculate the mean absolute differences (MAD) for a pair of paths (i, j) that only differ in one construction choice made at a specific fork, f, of our decision tree. This set of matched paths for a combination of estimator e and portfolio type p that only differs in one fork f can be defined as

$$S_f^{e,p} = \{(i,j) | c_{i,n} = c_{j,n} \quad \forall n \in \{1,\dots,7\}, c_{i,f} \neq c_{j,f}\} \ , \tag{4}$$

where $p = (c_1, \dots, c_F)$ is a path that is defined as a vector of choices, c_f, from Figure 2. Therefore, each path corresponds to one portfolio based on estimator e and portfolio type p. Note that we include transaction costs as a construction choice in this case because it is calculated before

the performance measure and its treatment as a construction choice grants insights into how its inclusion/neglection affects portfolio performance.

We calculate the differences in performance measures over the entire sample for each of the matched paths in $S_f^{e,p}$. We also create an overall estimate across all matched paths, S_f^p, for each portfolio type across all risk estimators. In this case, we also include the risk estimator as a decision fork for which we calculate the MAD. The mean absolute difference for each fork f, portfolio type p, and estimator e can be formalized as

$$MAD_f^p = \frac{1}{|S_f^{e,p}|} \sum_{(i,j) \in S_f^{e,p}} |m_i^e - m_j^e| \ , \tag{5}$$

for a selected performance measure m, where $|S_f^{e,p}|$ denotes the number of pairs in the set. With the MAD, we can quantify the impact of each construction choice on a selected performance measure. We report MADs across risk estimators and portfolio types. MAD results for Sharpe ratios are presented in Table 1 where decision forks are sorted from the largest overall impact (top) to the smallest overall impact (bottom) across all risk estimators. Further results for mean return, standard deviation, alpha, and maximum drawdown are available upon request.

From Table 1, we identify the portfolio construction decisions that induce the largest variation in Sharpe ratios when all other decisions at the respective fork are held constant.[9] Generally, the ranking among risk estimators and portfolio types is relatively consistent.

The most influential decision fork is the weighting scheme, followed by the consideration of transaction costs and the choice of the risk estimator. Across all risk estimators (Overall), the MAD impact of equally-weighting versus value-weighting on the SR ranges from 0.15 (*100-0*) to 0.20 (*130-30*). In comparison, MAD impact of transaction cost on the SR ranges from 0.06 (*100-0*) to 0.15 (*130-30*), and the choice of risk estimator from 0.07 (*100-0*) to 0.10 (*130-30*). The smallest impact arises from the price exclusion fork. Additionally, the length of the lookback window, rebalancing frequency, number of portfolio buckets, and size restrictions induce considerable and quantitatively similar variations in Sharpe ratios.

Consistent with the nonstandard error analysis in Figure 5, we find that portfolio construction choices have the most significant impact on *ewma* and *beta* portfolios, while idiosyncratic volatility portfolios (*ff3* and *ff5*) are the least affected. Finally, we observe larger (aggregated) MAD numbers for dollar-neutral portfolios, followed by fully-invested limited long-short portfolios and long-only portfolios. This provides further evidence that nonstandard errors and thus the impact of decision forks become more pronounced as the degrees of freedom in portfolio construction increase.[10]

Based on the reported MADs across the portfolio construction choices, the consideration of the risk estimator (estimation methodology and lookback window length) and transaction costs in our study

[9] MAD results for long-short portfolios are presented in Table A.1 in the Appendix.

[10] In unreported results we find a similar relationship between the impact of portfolio construction choices and other performance measures.

Table 1: Mean absolute differences at decision forks with respect to Sharpe ratio

This table presents the average mean absolute Sharpe ratio differences p.a., as defined in Equation 5, for each decision fork. For each fork, we compare Sharpe ratio differences between matched decision paths that vary only in the specific decision under consideration. These mean absolute differences are calculated individually for each risk estimator and averaged across all matched paths, as well as jointly for all risk estimators. The columns report averages for all risk estimators combined (Overall) and for individual risk estimators separately. The forks (rows) are ordered by their overall impact in descending order. The analysis is carried out separately for long-only (*100-0*) and limited-short (*130-30*) portfolios. Note that the impact of the risk estimator can only be calculated for the joint analysis across all estimators.

Portfolio	Fork	Overall	Hist.	EWMA	FF3	FF5	Beta
100-0	Weighting	**0.15**	0.17	0.17	0.14	0.13	0.14
	Risk Estimator	**0.07**					
	Trans. Costs	**0.06**	0.04	0.07	0.03	0.03	0.17
	Lookback	**0.03**	0.02	0.02	0.02	0.02	0.06
	Rebalancing	**0.02**	0.02	0.03	0.01	0.01	0.06
	Port. Buckets	**0.02**	0.02	0.02	0.01	0.01	0.05
	Drop Size	**0.02**	0.01	0.01	0.01	0.01	0.06
	Drop Price	**0.01**	0.00	0.00	0.00	0.00	0.02
130-30	Weighting	**0.20**	0.21	0.23	0.15	0.15	0.26
	Trans. Costs	**0.15**	0.11	0.18	0.09	0.09	0.28
	Risk Estimator	**0.10**					
	Drop Size	**0.06**	0.05	0.06	0.05	0.05	0.09
	Lookback	**0.06**	0.05	0.06	0.05	0.05	0.07
	Rebalancing	**0.06**	0.04	0.07	0.04	0.04	0.09
	Port. Buckets	**0.05**	0.03	0.04	0.03	0.03	0.11
	Drop Price	**0.02**	0.02	0.03	0.02	0.02	0.03

are, besides the weighting scheme, the most influential decisions. While the impact of the weighting scheme is generally known and already controlled for in many studies, these newly added decision forks are (practically) more relevant than the construction choices already covered by Hasler (2023), Walter et al. (2023), and Soebhag et al. (2024). Given the different order of decision forks across the performance metrics we investigate, portfolio managers should carefully evaluate their options to optimize the performance metric(s) that align most closely with their objectives. Focusing solely on average returns is insufficient to capture the complete picture.

3.3 Transaction Costs

A topic often overlooked in the academic literature on (factor) portfolio selection and nonstandard error analysis is the careful consideration of transaction costs. As shown in Table 1, the impact of including transaction costs in portfolio evaluation is as significant as the decision to invest using an equally-weighted or value-weighted approach. Therefore, an important contribution of our paper is the examination of the robustness of low-risk investing under realistic and practically relevant transaction cost considerations. To achieve this, we do not assume fixed transaction costs across stocks and time. Instead, we approximate transaction costs using half spreads, following the

methodology of Novy-Marx and Velikov (2016).

We first examine which portfolios have high turnover to see what characteristics lead to higher transaction costs due to more extensive reallocations of portfolios. In Figure 6, we plot the median yearly turnover for portfolios split by rebalancing frequencies and risk estimators as these are arguably the most influential choices on turnover. While the overall median turnover across all portfolios is 190%, the median turnover varies significantly by rebalancing frequency: 347% for weekly rebalancing, 204% for monthly rebalancing, 123% for quarterly rebalancing, and 65% for yearly rebalancing. For each rebalancing frequency, a consistent pattern emerges: Total and idiosyncratic volatility portfolios exhibit greater robustness over time, resulting in lower turnover compared to *ewma* and *beta* portfolios. This difference is a primary reason why net-of-transaction-cost results are less favorable for *ewma* and *beta* portfolios. The higher turnover for *ewma* portfolios arises because these portfolios place greater weight on recent observations to improve volatility estimation. However, this approach has the drawback of being more erratic and susceptible to influence from a few observations, depending on the chosen decay factor, λ. Instead, the turnover for *beta* portfolios is elevated due to the instability of OLS regression coefficient estimates.

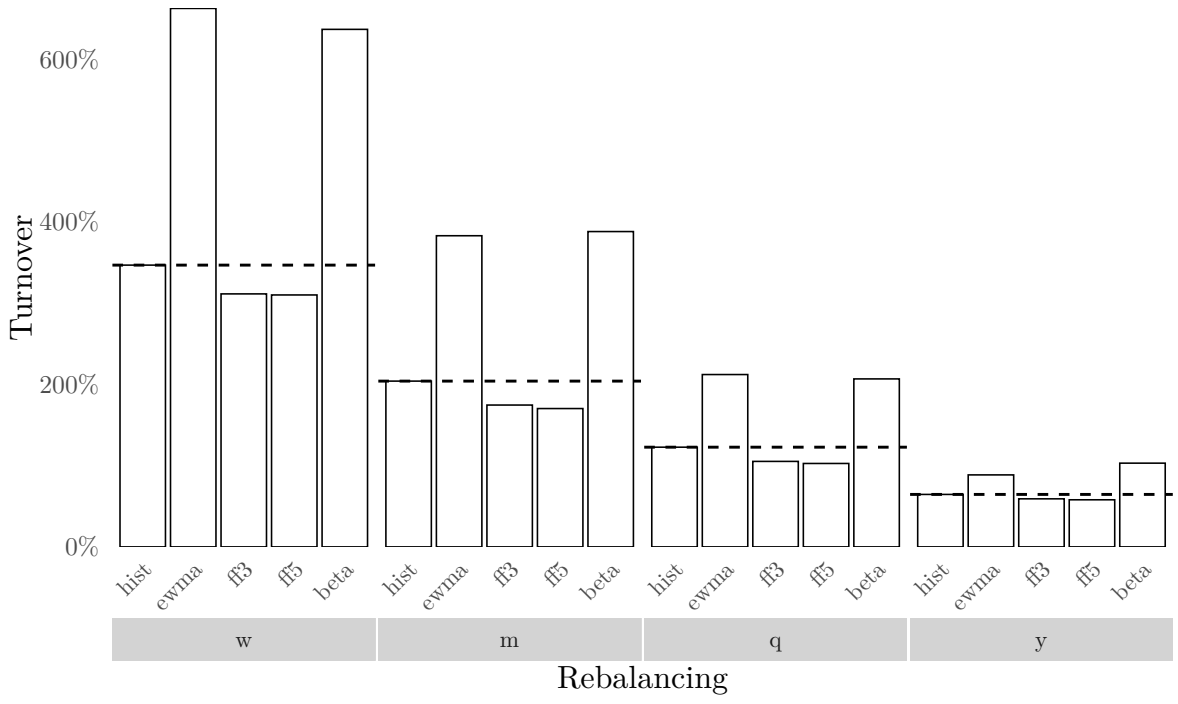

Figure 6: Median yearly portfolio turnover

The plot illustrates median yearly turnover for portfolios across construction choices. Turnover is measured for portfolios grouped by the same risk estimator and rebalancing frequency: weekly (w), monthly (m), quarterly (q), and yearly (y). The median yearly turnover across all portfolios is 190%. The turnover is calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

Next, we investigate what characteristics lead to higher transaction costs due to putting more weight on stocks that are more costly to trade. In Figure 7, we analyze the median transaction costs

(per unit of turnover) and split our results by portfolio type, portfolio weighting, and size filters. We perform this "size focused" split because, like many other anomalies, the low-risk anomaly is criticized for shorting expensive small stocks, which results in high transaction costs that are often overlooked (Baker et al., 2011; Novy-Marx & Velikov, 2022). We find that the overall median transaction cost per unit of turnover across all portfolios is 0.40%. However, it varies by portfolio type: 0.35% for long-only portfolios and 0.44% for limited long-short portfolios. This shows that, as criticized by the literature, the short leg of low-volatility portfolios is indeed more costly to trade. As expected, we observe lower costs per turnover for value-weighted portfolios compared to equally-weighted ones, as well as for portfolios that exclude small stocks. Both the weighting methodology and the size filter influence exposure to expensive small stocks, directly impacting the costs per unit of turnover for the respective portfolios. Notably, however, only equally weighted portfolios without a size filter exhibit exceptionally high trading costs, while other specifications show comparatively small costs. These findings highlight why "Drop Size" is one of the influential portfolio construction choices discussed in Section 3.2 and should not be underestimated.

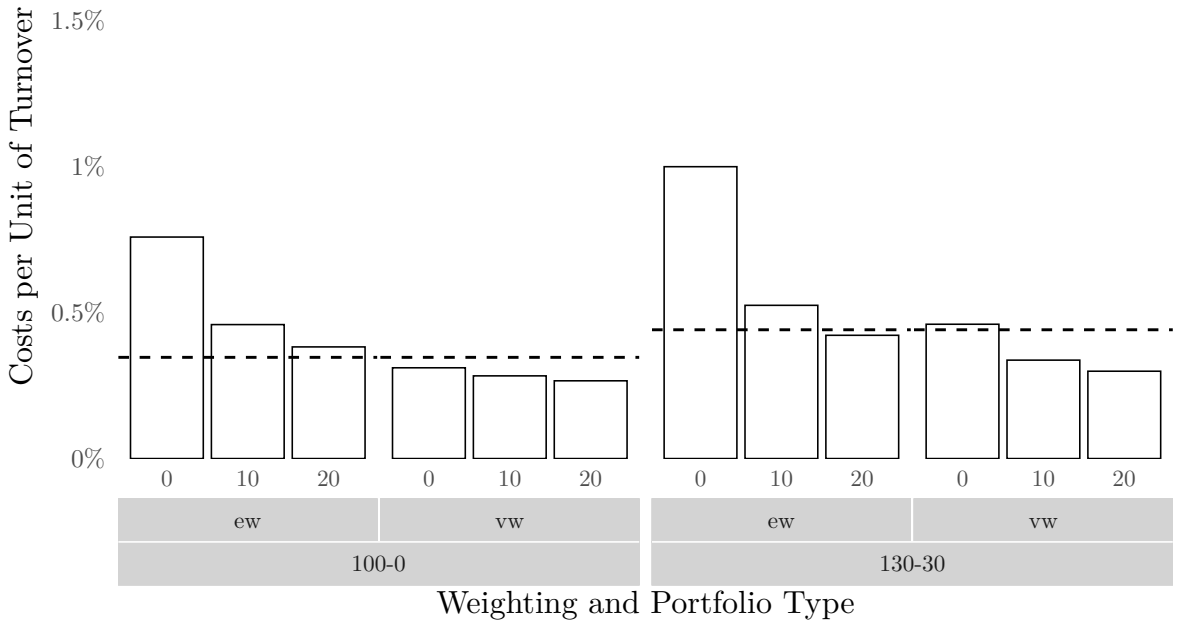

Figure 7: Median transaction costs per unit of turnover for portfolios with varying portfolio type, weighting scheme, and size filters

The plot illustrates median costs per unit of turnover for portfolios across construction choices, categorized by three decision variables: size exclusion, weighting scheme, and portfolio type. Size exclusion refers to the exclusion of stocks that are smaller than the 0, 10, or 20 percentile of stocks with respect to market equity. Weighting scheme indicates whether portfolios are constructed using value-weighing (vw) or equal-weighing (ew). Considered portfolio types are long-only (*100-0*) and limited-short (*130-30*) portfolios. The median costs per turnover across all portfolios within a portfolio type is indicated by the dashed black lines. The median transaction costs per unit of turnover across all portfolios is 0.40%. The costs are calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

From our analysis, we conclude that portfolio construction choices significantly impact the turnover and average stock-level costs of low-risk portfolios that ultimately determine total transaction costs. We argue that for practical applications, transaction costs must be properly accounted for, as they can significantly influence inferences about what portfolio construction choices lead to positive performance. Our findings show that transaction costs influence, most notably, the selection of the optimal risk estimator, the use of size filters, and the preferred rebalancing frequency.

3.4 Determinants of Outperformance

In this section, we examine how frequently the considered portfolios outperform the benchmark and identify the main drivers of any outperformance. First, we report the percentage of portfolios that exceed the benchmark across multiple performance metrics in Table 2.

Table 2: Percentage of portfolios that outperform their benchmark with and without costs

The table reports the percentage of portfolios within each portfolio type that outperform their respective benchmarks based on seven performance metrics: cumulative return, average return, alpha, generalized alpha, standard deviation, Sharpe ratio, and maximum drawdown. The results are further divided by whether transaction costs are included in the evaluation. The benchmark is a value-weighted market portfolio. The statistics are calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

Portfolio	Trans. Costs	Cum.	Ret.	Alpha	Alpha$_g$	Std.	SR	m. DD
100-0	w/o	81%	65%	42%	100%	100%	100%	97%
100-0	with	64%	44%	18%	99%	100%	94%	96%
130-30	w/o	77%	68%	44%	100%	100%	94%	84%
130-30	with	45%	31%	12%	94%	100%	72%	71%

In general, we observe that the percentage of portfolios outperforming the benchmark is higher for long-only portfolios across performance measures and transaction cost considerations. This is primarily due to the reduced impact of nonstandard errors, as discussed in Section 3.2. Specifically, around 81% of the gross cumulative portfolio returns are higher for long-only portfolios compared to the value-weighted benchmark, while 64% of net cumulative returns outperform. However, these percentages are noticeably smaller when comparing annualized arithmetic mean returns: 65% of long-only portfolios outperform the benchmark in terms of gross arithmetic mean returns, but only 44% outperform in terms of net arithmetic mean returns. This difference can be attributed to compounding effects, where higher (lower) volatility leads to lower (higher) geometric average returns.[11]

The next two columns of Table 2 show the percentage of portfolios with positive alpha and positive generalized alpha from the daily Fama and French (2015) five-factor regression. For gross returns,

[11] This phenomenon, which disproportionately affects the most volatile stocks and benchmarks, is further discussed by van Vliet et al. (2011) and Spitznagel (2021). Arithmetic averages ignore compounding effects, which are particularly important when comparing portfolios with significantly different volatility characteristics, as is evident in our empirical analysis. Arguably, geometric (compounded) average returns are more relevant to long-term investors than arithmetic (simple) averages.

less than half of the portfolios exhibit positive alpha, and for net returns, less than a fifth do. However, almost all portfolios exhibit a positive generalized alpha. Novy-Marx (2015) argues that the low-beta and low-volatility anomalies can be explained by a three-factor model augmented with a profitability factor. Similarly, Fama and French (2016) find that their five-factor model, which adds profitability and investment factors to the original three-factor model, explains returns on low-risk-sorted portfolios. Nevertheless, Blitz and Vidojevic (2017) and related work contend that the low-risk anomaly is not fully explained by the five-factor model: *"But as long as the data indicates that portfolios with higher risk do not generate higher returns, it is premature to conclude that the low-risk anomaly has been resolved."* Consistent with this view, our results confirm that residual performance in low-risk–sorted portfolios remains that is not fully captured by the five-factor model.

The last three columns of Table 2 present risk and risk-adjusted performance metrics. As we focus on low-risk portfolios, it comes as no surprise that risk metrics improve relative to the benchmark. Nevertheless, it is noteworthy that the standard deviation of all portfolios consistently remains lower than the benchmark, while Sharpe ratios and maximum drawdowns improve across nearly all portfolio constructions. Results are especially robust for the long-only portfolios: roughly 100% of gross Sharpe ratios exceed the benchmark and 97% of gross maximum drawdowns are lower. On a net basis, 94% of Sharpe ratios are higher and 96% of maximum drawdowns are lower than the benchmark.

Across all 6,480 portfolio variations, the average Sharpe ratio is 0.57, which exceeds the benchmark Sharpe ratio of 0.5. In Figure 8, we illustrate the distribution of these Sharpe ratios, categorized by various construction choices.

The results can be summarized as follows:

- **Risk Estimator:** The total volatility estimator performs the best, followed by *ewma* and idiosyncratic volatility estimators. Portfolios based on *beta* sorting perform the worst by far and also have the largest dispersion.
- **Portfolio Type:** Interestingly, long-only (*100-0*) portfolios outperform limited-short (*130-30*) portfolios and exhibit lower dispersion.
- **Size Exclusion:** The inclusion of size filters boosts performance.
- **Price Exclusion:** The impact of price filters is less pronounced, but higher price filters are preferred.
- **Lookback Window:** There is no obvious pattern regarding length of the lookback window. A shorter lookback of one year provides a more attractive distribution by increasing Q1.
- **Rebalancing Frequency:** Less frequent rebalancing (quarterly or annual) outperforms weekly and monthly updates.
- **Portfolio Buckets:** A smaller number of buckets, resulting in a larger number of stocks per portfolio, outperforms.

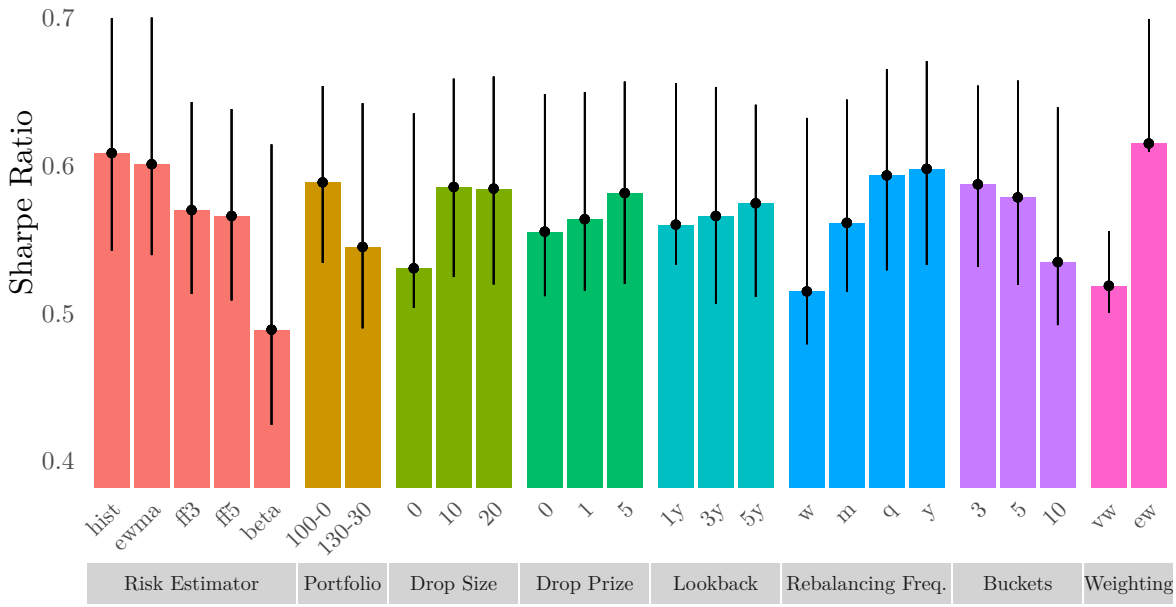

Figure 8: Average Sharpe ratios by decision variable including transaction costs

The plot displays the average annualized Sharpe ratios of long-only and limited-short portfolios after accounting for transaction costs. Each bar represents portfolios sharing a specific decision variable within a given decision fork, differentiated by colors. The black lines show the inter quartile range of Sharpe ratios. Sharpe ratios are calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

- **Weighting Scheme:** Equally-weighted portfolios substantially outperform value-weighted portfolios, though they also exhibit markedly larger dispersion.

Note that, in unreported results, we assess which portfolio construction choices are most suitable for benchmarked managers by replicating the analysis with information ratio (IR) as a performance measure. We find that the information ratios distributions conditional to construction choices closely mirror the Sharpe ratio distributions above.

In Appendix B, we complement the main results with a "winners–losers" audit of implementable portfolios, net of transaction costs. We tally how often each construction choice appears among the top-100 winners and bottom-100 losers across multiple performance metrics, highlighting statistically notable over/under-representation (green/red) in Table B.1, and we document the design choices of the top and bottom Sharpe portfolios in Table B.2. Three patterns stand out. First, the weighting scheme is the dominant lever: equal- versus value-weighting drives the largest shifts in outcomes. Second, risk estimators matter next – total-volatility measures (*hist*, *ewma*) tend to deliver higher Sharpe ratios than *beta*, which lowers volatility but at the cost of mean returns. Third, lower-turnover choices (coarser rebalancing), basic filters (price/size), and larger buckets generally improve net performance and temper drawdowns, whereas more aggressive settings mainly widen

dispersion. These conclusions align with those from the Sharpe ratio analysis above and remain stable over time and during recessions, as shown in Appendix C.

3.5 A Recommendation for Nonstandard Errors in Low-Risk Portfolios

As the final step in our analysis, we aim to recommend an attractive and robust construction methodology for low-risk portfolios. Synthesizing our results, the most natural baseline for a low-risk portfolio consistent with Figure 8, is a long-only specification that estimates risk using historical volatility, applies a size filter at the 10th percentile and a price filter of five USD, uses a five-year lookback window, and forms an equally-weighted tercile portfolio that is rebalanced yearly. This configuration balances net performance and robustness, however, the *optimal* portfolio is derived from an in-sample manner.

Additionally, we aim to not induce any look-ahead bias from the insights of our previous analyses and thus construct a feasible low-risk portfolio with information that was available to investors at that given point in time, thus leading to a portfolio that could have been implemented by investors in the past. Specifically, beginning in 1979, at the end of each calendar year, we evaluate the performance of all low-risk portfolios by calculating their Sharpe ratios using all historical data available up to that point. Since earlier results have shown that certain specifications can lead to extreme outliers, we construct an *ensemble* strategy that selects the 100 portfolios with the highest Sharpe ratios and invests in their equal-weighted average for the subsequent year. This procedure is designed to identify portfolios that deliver consistent outperformance while reducing the influence of individual outliers. To ensure that the evaluation reflects realistic investment scenarios, we limit our focus to portfolios that account for transaction costs.

The cumulative returns of the optimal portfolio (blue), the ensemble portfolio (red), and the benchmark (black) are illustrated in Figure 9. Both the optimal and ensemble portfolios significantly outperform the benchmark over the sample period, with much of the outperformance driven by the dot-com bubble in the early 2000s. Notably, while low-risk strategies underperformed during the buildup of the bubble, they recovered strongly thereafter.

The outperformance of the optimal and ensemble strategy is further examined in Table 3, which presents selected performance metrics. The results indicate that both low-risk portfolios consistently outperform or perform at least on par with the benchmark across all key performance metrics. They achieve at least a similar average return and a lower standard deviation, resulting in markedly higher Sharpe ratio. Additionally, the low-risk portfolios exhibit a smaller maximum drawdown compared to the benchmark, underscoring its resilience during market downturns. We further eval-

[12] To test the outperformance of the ensemble portfolio over the benchmark a two-sided p-value for the null hypothesis of equal standard deviations, respectively equal Sharpe ratios, is obtained by the prewhitened HAC_{PW} method described in Ledoit and Wolf (2011, Section 3.1), respectively Ledoit and Wolf (2008, Section 3.1). As the out-of-sample size is very large at 11,598, there is no need to use the computationally more involved bootstrap method described in Ledoit and Wolf (2011, Section 3.2), respectively in Ledoit and Wolf (2008, Section 3.2), which is preferred for small sample sizes.

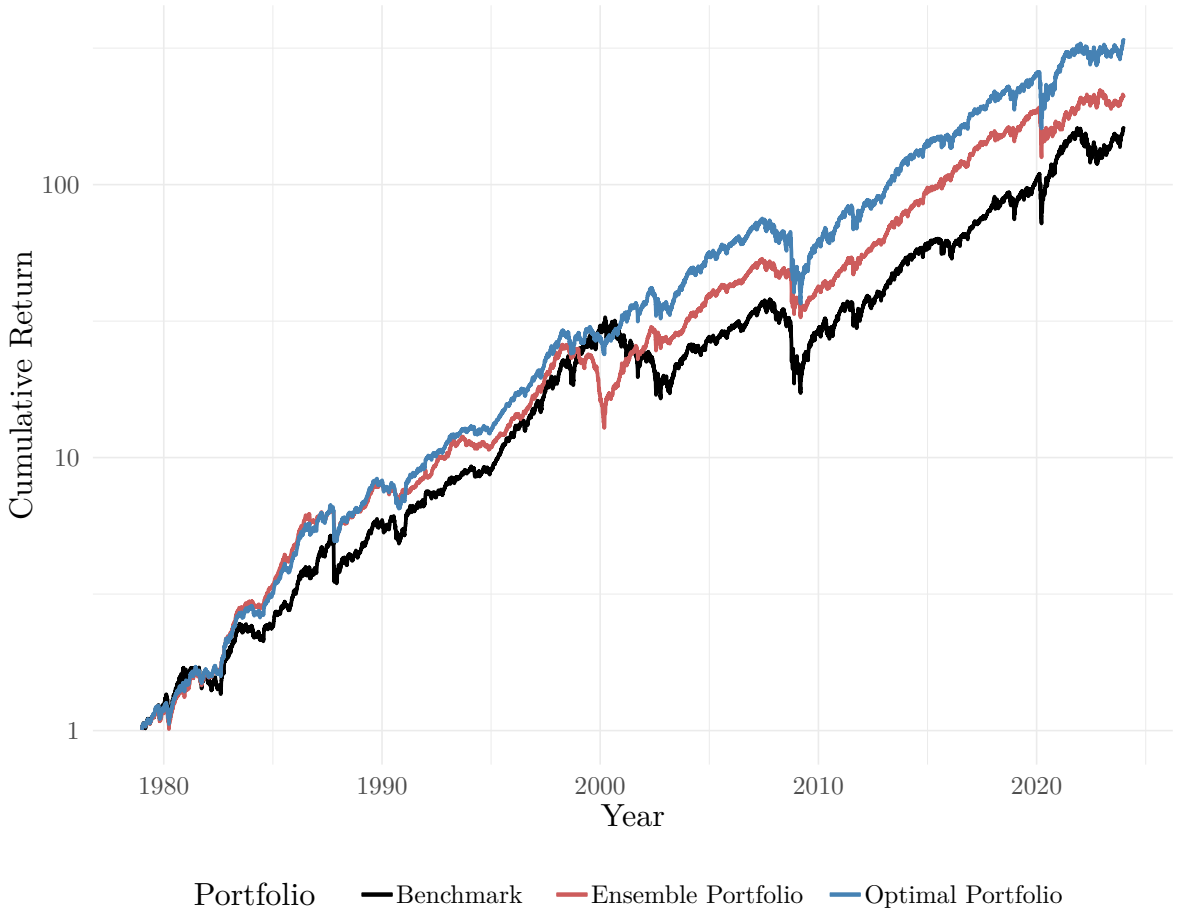

Figure 9: Cumulative returns of the benchmark and the recommended low-risk portfolios

The figure illustrates cumulative returns for our rolling ensemble portfolio (red), our optimal Sharpe ratio portfolio (blue), alongside a value-weighted market benchmark (black). The ensemble portfolio is constructed at the beginning of each calendar year by selecting the 100 low-risk portfolios with the highest Sharpe ratios, evaluated using an expanding window of daily, net-of-transaction-costs returns available up to that point. The portfolio then invests equally in this set of strategies for the subsequent year. The optimal is the portfolio constructed from the decision choices that each create the highest Sharpe ratio in their respective decision fork. The cumulative returns are calculated based on 11,346 daily out-of-sample returns from January 1979 to December 2023 and presented on a log-10 y-axis.

uate whether the outperformance of the low-risk portfolios in terms of lower standard deviation and higher Sharpe ratio compared to the benchmark is statistically significant and economically meaningful. For standard deviation (Std.) and Sharpe ratio (SR) the $*$, respecitvely $**$, indicate that these differences are statistically significant at the 5%, respectively 1%, significance level.[12]

Lastly, we examine the types of portfolios selected by the ensemble strategy. To this end, Table D.1 reports the relative frequency with which different portfolio construction choices appear in the ensemble portfolio. We find that portfolios based on idiosyncratic volatility (*ff3* and *ff5*) are rarely selected, whereas other risk estimators are more evenly represented. In terms of portfolio type, the ensemble strategy tends to favor limited-short portfolios. Both size and price filters appear more

Table 3: Summary statistics of recommended low-risk portfolios vs. benchmark

The table holds portfolio summary statistics for our rolling ensemble portfolio and our optimal portfolio alongside a value-weighted market benchmark. The standard deviation and Sharpe ratio of the ensemble and optimal portfolios are assessed relative to those of the benchmark portfolio. Statistically significant differences at the 1% and 5% levels are denoted by ** and *, respectively. The performance metrics are calculated based on 11,346 daily out-of-sample returns from January 1979 to December 2023.

Portfolio	Cum.	Ret.	Std.	SR	IR	m. DD
Ensemble	210.8	12.6%	11.4%**	0.75*	−0.02	50.4%
Optimal	336.4	14.0%	14.7%**	0.68**	0.17	51.3%
Benchmark	172.6	12.7%	17.4%	0.50		54.7%

frequently among selected portfolios, suggesting that these filters generally enhance performance. For the lookback window, shorter horizons are preferred, indicating that more recent information contributes more to performance than longer, more stable estimation periods. Rebalancing frequency also shows a clear pattern: portfolios with less frequent rebalancing are selected more often, again indicating that the costs of more frequent portfolio adjustments outweigh their potential benefits. No dominant trend emerges for portfolio buckets, which are relatively evenly represented. Finally, nearly all selected portfolios are equally weighted, consistent with our earlier findings that equally-weighted portfolios tend to deliver better performance metrics.

Overall, this section demonstrates that, when managed effectively, low-risk strategies can deliver a favorable risk-return profile relative to the market. Furthermore, it highlights that while nonstandard errors can significantly influence results for low-risk portfolios, the best-performing portfolios are constructed in a risk-controlled manner – employing size and price filters along with infrequent rebalancing. This approach reduces result dispersion and enables consistent outperformance over time.

4 Conclusion

This study provides a comprehensive examination of the impact of portfolio construction choices on the performance of U.S. low-risk equity portfolios. Our analysis highlights that the methodological decisions during the portfolio construction process significantly influence performance metrics and are crucial for understanding and enhancing investment outcomes.

This paper makes several significant contributions to the literature. First, it systematically evaluates the influence of risk estimators on low-risk portfolio performance, demonstrating that the choice of estimation methodology – specifically, the risk estimator choice together with its lookback window length – accounts for a big share of performance variability. Second, the study investigates the role of transaction costs and performance evaluation metrics, shedding light on the trade-offs between return enhancement and cost efficiency – areas that have been largely neglected in the current literature. We demonstrate that these enhancements are essential, as both significantly shape

preferences and inferences regarding portfolio construction choices. Third, our work contributes by focusing on realistic, implementable portfolios that address practical challenges such as transaction costs and leverage constraints. By bridging the gap between theory and practice, it offers valuable insights for both academics and practitioners.

Our results show that nonstandard errors are generally smaller for long-only investors, providing encouraging evidence that low-risk investing is more difficult to misapply. This robustness strengthens the case for low-risk strategies in practical applications. At the same time, our analysis highlights that methodological choices matter: the decision between equal-weighting and value-weighting proves to be the most consequential, followed by the choice of risk estimator. The Sharpe ratio results further reinforce these conclusions, showing consistent patterns across specifications: total volatility performs best, beta-sorting worst; long-only strategies outperform 130/30 implementations; size filters add value; less frequent rebalancing proves advantageous; and equal-weighting generally delivers superior outcomes. By bringing these insights to the foreground, we emphasize both the resilience of low-risk investing and the importance of carefully considering portfolio construction choices when implementing such strategies.

Overall, this study highlights the dispersion in decision choices for sorting variable computation and portfolio construction. It underscores the risk of nonstandard errors while demonstrating how construction choices can be optimized to maximize the benefits of the low-risk anomaly. Our results indicate that optimal low-risk portfolios should also be constructed in a low-risk manner. This means that choices that arguably reduce risk, such as the exclusion of small and penny stocks, long rebalancing cycles, and little short selling, lead to consistent performance over time that is not driven by single events.

References

Abdi, F., & Ranaldo, A. (2017). A simple estimation of bid-Ask spreads from daily close, high, and low prices. *Review of Financial Studies*, *30*(12), 4437–4480.

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.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2009). High idiosyncratic volatility and low returns: International and further U.S. evidence. *Journal of Financial Economics*, *91*(1), 1–23.

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2014). Low-risk investing without industry bets. *Financial Analysts Journal*, *70*(4), 24–41.

Baker, M., Bradley, B., & Wurgler, J. (2011). Benchmarks as Limits to Arbitrage: Understanding the Low Volatility Anomaly. *Financial Analysts Journal*, *67*(1), 40–54.

Baker, N. L., & Haugen, R. A. (2012). Low Risk Stocks Outperform within All Observable Markets of the World. *SSRN Electronic Journal*.

Blitz, D., Howard, C., Huang, D., & Jansen, M. (2024). Low-Risk Alpha Without Low Beta. *SSRN Electronic Journal*.

Blitz, D., Pang, J., & van Vliet, P. (2013). The volatility effect in emerging markets. *Emerging Markets Review*, *16*(1), 31–45.

Blitz, D., van Vliet, P., & Baltussen, G. (2023). The Volatility Effect Revisited. *The Journal of Portfolio Management*, *46*(2), 45–63.

Blitz, D., & van Vliet, P. (2007). The volatility effect. *The Journal of Portfolio Management*, *34*(1), 102–113.

Blitz, D., van Vliet, P., & Baltussen, G. (2020). The volatility effect revisited. *Journal of Portfolio Management*, *46*(2), 45–63.

Blitz, D., & Vidojevic, M. (2017). The profitability of low-volatility. *Journal of Empirical Finance*, *43*, 33–42.

Bryzgalova, S., Lerner, S., Lettau, M., & Pelger, M. (2024). Missing Financial Data. *The Review of Financial Studies*, *00*, 1–80.

Chen, M., Hanauer, M. X., & Kalsbach, T. (2024). Design choices, machine learning, and the cross-section of stock returns. *SSRN Electronic Journal*.

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

De Nard, G. (2022). Oops! I shrunk the sample covariance matrix again: Blockbuster meets shrinkage. *Journal of Financial Econometrics*, *20(4)*, 569–611.

Detzel, A., Novy-Marx, R., & Velikov, M. (2023). Model Comparison with Transaction Costs. *Journal of Finance*, *78*(3), 1227–1831.

Dimson, E., Marsh, P., & Staunton, M. (2017). Factor-Based Investing: The Long-Term Evidence. *The Journal of Portfolio Management*, *43*(5), 15–37.

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

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

Fama, E. F., & French, K. R. (2016). Dissecting Anomalies with a Five-Factor Model. *Review of Financial Studies*, *29*(1), 69–103.

Fama, E. F., & MacBeth, J. (1973). Risk, Return and Equilibrium: Empirical Tests. *Journal of Political Economy*, *81*(3), 697–637.

Frazzini, A., & Pedersen, L. H. (2014). Betting Against Beta. *Journal of Financial Economics*, *111*(1), 1–25.

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

Hasler, M. (2023). Looking Under the Hood of Data-Mining. *SSRN Electronic Journal*.

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

Invesco. (2023). *2023 Invesco Global Systematic Investing Study*.

Invesco. (2024). *2024 Invesco Global Systematic Investing Study*.

Jagannathan, R., & Ma, T. (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *Journal of Finance*, *54*(4), 1651–1684.

Jensen, T. I., Kelly, B. T., & Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance*, *78*(5), 2465–2518.

Ledoit, O., & Wolf, M. (2011). Robust performance hypothesis testing with the variance. *Wilmott Magazine*, *September*, 86–89.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, *15*(5), 850–859.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neus, S., Razen, M., & Weitzel, U. (2024). Non-Standard Errors *. *The Journal of Finance*, *79*(3), 2339–2390.

Mina, J., & Xiao, J. Y. (2001). Return to RiskMetrics: The Evolution of a Standard.

Novy-Marx, R. (2015). How Can a Q-Theoretic Model Price Momentum? *NBER Working Paper Series*, 20985.

Novy-Marx, R., & Velikov, M. (2016). A Taxonomy of Anomalies and Their Trading Costs. *Review of Financial Studies*, *29*(1), 104–147.

Novy-Marx, R., & Velikov, M. (2022). Betting against betting against beta. *Journal of Financial Economics*, *143*(1), 80–106.

RiskMetrics. (1996). *RiskMetricsTM—Technical Document*.

Soebhag, A., Baltussen, G., & van Vliet, P. (2025). Factoring in the Low-Volatility Factor. SSRN Working Paper.

Soebhag, A., Van Vliet, B., & Verwijmeren, P. (2024). Non-Standard Errors in Asset Pricing: Mind Your Sorts. *Journal of Empirical Finance*, *78*, 101517.

Spitznagel, M. (2021). *Safe Haven: Investing for Financial Storms*. Wiley.

Traut, J. (2023). What we know about the low-risk anomaly: a literature review. *Financial Markets and Portfolio Management*, *37*(3), 297–324.

van Vliet, P., Blitz, D., & van der Grient, B. (2011). Is the relation between volatility and expected stock returns positive, flat or negative? *SSRN working paper of Robeco*.

Walter, D., Weber, R., & Weiss, P. (2023). Methodological uncertainty in portfolio sorts. *SSRN Electronic Journal*.

Zumbach, G. O. (2007). The Riskmetrics 2006 Methodology. *SSRN Electronic Journal*.

Appendices

A Dollar-Neutral Long-Short Portfolios

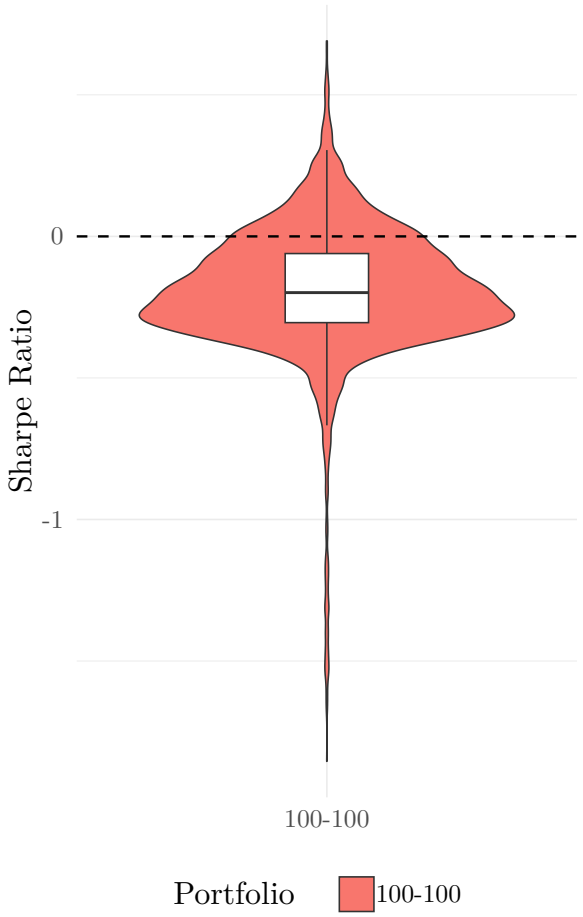

Figure A.1: Sharpe ratios of long-short portfolios across portfolio construction choices

This figure plots the distribution of (annualized) gross (without transaction costs) and net (with transaction costs) Sharpe ratios for all the 3,240 long-short portfolios across varying portfolio construction choices. The benchmark Sharpe ratio of zero is shown as a dashed line. The box represent the interquartile range and the median of the distribution. All the portfolios are based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

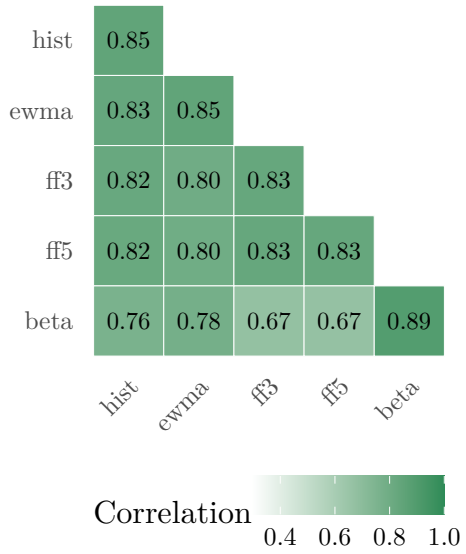

Figure A.2: Average correlations for long-short portfolios between and within an estimator class including transaction costs

This figure shows the average correlation between two portfolios, of which one is based on an estimator belonging to the class on the corresponding row, while the other one is based on an estimator from the class on the corresponding column. In particular, the diagonal elements are average correlations between two portfolios based on estimators within a certain class. The analysis is carried out separately for long-short (*100-100*) portfolios. Correlations are calculated based on 11,598 daily out-of-sample returns including transaction costs from January 1978 to December 2023.

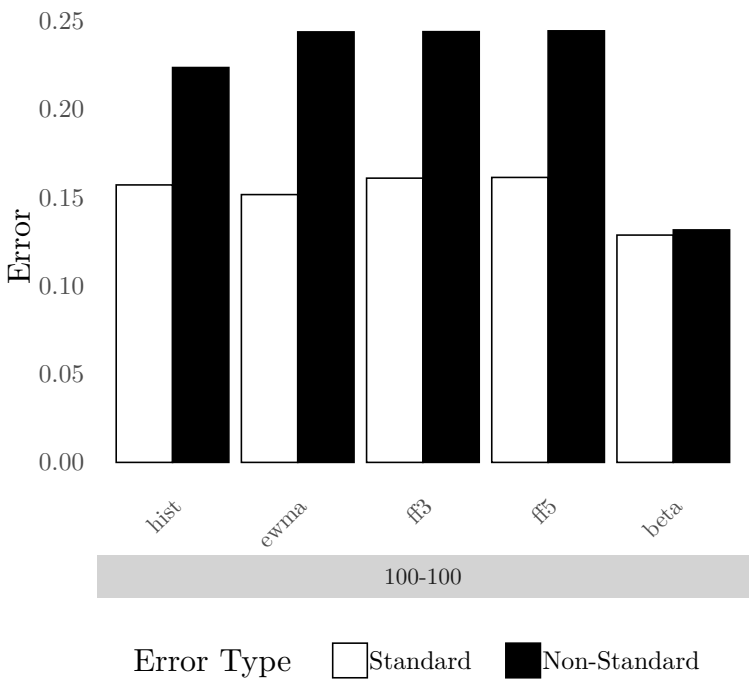

Figure A.3: Standard vs. nonstandard errors for long-short portfolios

This figure plots the nonstandard error (black) and standard error (white) for each risk estimator of long-short portfolios. In particular, for a given risk estimator, we define the nonstandard error as the standard deviation of the Sharpe ratios across all portfolios calculated using the specified risk estimator. In contrast, the standard error is defined by first computing the time series of monthly Sharpe ratios for each portfolio and then calculating its standard deviation. The standard error for a risk estimator is the average of these standard deviations across all portfolios based on the given risk estimator. Both error types are annualized and calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

Table A.1: Mean absolute differences at decision forks with respect to Sharpe ratio for long-short portfolios

This table presents the average mean absolute Sharpe ratio differences p.a., as defined in Equation 5, for each decision fork. For each fork, we compare Sharpe ratio differences between matched decision paths that vary only in the specific decision under consideration. These mean absolute differences are calculated individually for each risk estimator and averaged across all matched paths, as well as jointly for all risk estimators. The columns report averages for all risk estimators combined (Overall) and for individual risk estimators separately. The forks (rows) are ordered by their overall impact in descending order. The analysis is carried out for long-short (*100-100*) portfolios. Note that the impact of the risk estimator can only be calculated for the joint analysis across all estimators.

Portfolio	Fork	Overall	Hist.	EWMA	FF3	FF5	Beta
100-100	Trans. Costs	**0.17**	0.16	0.20	0.18	0.18	0.15
	Weighting	**0.12**	0.12	0.13	0.13	0.13	0.09
	Risk Estimator	**0.10**					
	Port. Buckets	**0.10**	0.11	0.11	0.12	0.12	0.05
	Drop Size	**0.10**	0.11	0.11	0.12	0.12	0.04
	Lookback	**0.09**	0.09	0.09	0.11	0.11	0.04
	Rebalancing	**0.08**	0.08	0.09	0.08	0.09	0.04
	Drop Price	**0.03**	0.04	0.04	0.04	0.04	0.01

B Winner Loser Evaluation

It is natural to ask which implementable portfolio construction choices perform the best and worst with respect to various performance measures, net of transaction costs. To address this question, we report the relative counts of decision variables of the 100 winner (W) and 100 loser (L) portfolios in Table B.1. We highlight results in (light) green if they are more than (two) three standard deviations away from an equal distribution of variables in a positive direction. For winning portfolios, this indicates that the variable is overrepresented, while for loser portfolios, it suggests the variable is underrepresented. Conversely, we apply (light) red coloring for the opposite scenario. We further report the construction choices of the top and bottom five Sharpe ratio portfolios across different portfolio types and transaction cost considerations in Table B.2.

Table B.1: Relative appearance of portfolio construction choices in top 100 winner (W) and loser (L) portfolios for long-only and limited-short portfolios with transaction costs

The table presents the relative occurrence of decision variables in the top 100 winner (W) and bottom 100 loser (L) portfolios, ranked according to various performance metrics. These portfolios, encompassing both long-only and limited-short strategies, are evaluated net of transaction costs. The relative appearance of decision variables is calculated for each decision fork across the following performance metrics: average return, alpha, volatility, Sharpe ratio, information ratio, and maximum drawdown. Results are highlighted in (light) green if they are more than (two) three standard deviations away from an uniform distribution of variables in a positive direction. For winning portfolios, this indicates that the variable is overrepresented, while for loser portfolios, it suggests the variable is underrepresented. Conversely, we apply (light) red coloring for the opposite scenario. The performance metrics are calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

	Return		Alpha		Volatility		SR		IR		max. DD	
Variable	W	L	W	L	W	L	W	L	W	L	W	L
Risk Estimator												
hist	0%	4%	16%	12%	0%	0%	35%	5%	1%	7%	0%	8%
ewma	21%	11%	49%	19%	3%	0%	58%	11%	6%	12%	1%	19%
ff3	42%	3%	0%	16%	0%	50%	0%	5%	48%	9%	0%	7%
ff5	37%	3%	0%	17%	0%	50%	0%	5%	45%	8%	0%	7%
beta	0%	79%	35%	36%	97%	0%	7%	74%	0%	64%	99%	59%
Portfolio Type												
100-0	37%	20%	14%	9%	14%	5%	0%	19%	99%	33%	19%	15%
130-30	63%	80%	86%	91%	86%	95%	100%	81%	1%	67%	81%	85%
Size Exclusion												
0	42%	85%	47%	87%	94%	77%	21%	88%	34%	94%	48%	97%
10	47%	10%	32%	7%	5%	11%	42%	7%	42%	6%	28%	3%
20	11%	5%	21%	6%	1%	12%	37%	5%	24%	0%	24%	0%
Prize Exclusion												
0	36%	50%	33%	57%	18%	54%	25%	52%	44%	53%	30%	57%
1	36%	37%	26%	37%	33%	33%	26%	35%	38%	37%	36%	36%
5	28%	13%	41%	6%	49%	13%	49%	13%	18%	10%	34%	7%
Lookback Window												
1y	73%	64%	48%	58%	27%	20%	78%	67%	40%	60%	43%	60%
3y	18%	24%	26%	24%	35%	41%	14%	22%	48%	26%	22%	27%
5y	9%	12%	26%	18%	38%	39%	8%	11%	12%	14%	35%	13%
Rebalancing Frequency												
w	18%	60%	0%	84%	37%	12%	0%	63%	33%	64%	18%	56%
m	3%	32%	7%	16%	23%	12%	16%	29%	1%	30%	39%	24%
q	39%	8%	36%	0%	23%	18%	55%	8%	29%	6%	34%	8%
y	40%	0%	57%	0%	17%	58%	29%	0%	37%	0%	9%	12%
Portfolio Buckets												
3	73%	8%	9%	16%	25%	5%	31%	8%	100%	14%	66%	8%
5	9%	22%	23%	27%	40%	8%	45%	24%	0%	30%	15%	24%
10	18%	70%	68%	57%	35%	87%	24%	68%	0%	56%	19%	68%
Weighting Scheme												
vw	17%	12%	0%	14%	0%	86%	0%	12%	0%	1%	100%	10%
ew	83%	88%	100%	86%	100%	14%	100%	88%	100%	99%	0%	90%

Table B.2: Characteristics of top and bottom portfolios

This table shows the construction choices of the top five and bottom five portfolios in term of average Sharpe ratio. The analysis is conducted separately for each portfolio type: long-only (*100-0*) on top and limited-short (*130-30*) on the bottom as well as without transaction costs (left) and with transaction costs (right). The Sharpe ratio is calculated based on 11,598 daily out-of-sample returns from January 1978 to December 2023.

	Top 5					Bottom 5					Top 5					Bottom 5				
	Top	2	3	4	5	5	4	3	2	Bot.	Top	2	3	4	5	5	4	3	2	Bot.
	100-0 without transaction costs										100-0 with transaction costs									
SR	1.59	1.56	1.53	1.47	1.46	0.48	0.47	0.47	0.47	0.46	0.76	0.75	0.75	0.75	0.75	-0.17	-0.45	-0.54	-0.56	-0.59
Risk Estimator	beta	beta	beta	beta	beta	ff5	ff5	ff5	ff5	beta	beta	beta	ewma	hist	hist	beta	beta	beta	beta	beta
Drop Size	0	0	0	0	0	10	20	20	20	0	0	0	0	0	0	0	0	0	0	0
Drop Prize	0	0	0	1	0	5	5	1	0	0	5	5	5	5	1	0	1	1	0	0
Lookback	3y	5y	1y	5y	3y	3y	3y	3y	3y	1y	5y	5y	5y	3y	3y	1y	1y	1y	1y	1y
Rebalancing	w	w	w	w	w	y	y	y	y	y	y	y	w	w	w	m	m	w	m	w
Buckets	10	10	10	10	5	10	10	10	10	10	10	5	10	10	10	5	10	10	10	10
Weighting	ew	ew	ew	ew	ew	vw	vw	vw	vw	vw	ew	ew	ew	ew	ew	ew	ew	ew	ew	ew
	130-30 without transaction costs										130-30 with transaction costs									
SR	1.93	1.88	1.86	1.82	1.80	0.34	0.33	0.32	0.29	0.29	0.88	0.87	0.86	0.86	0.85	-0.70	-0.77	-0.86	-1.12	-1.19
Risk Estimator	beta	beta	beta	beta	beta	beta	beta	beta	beta	beta	ewma	ewma	ewma	ewma	ewma	beta	beta	beta	beta	beta
Drop Size	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
Drop Prize	0	0	0	0	0	5	1	1	0	0	5	5	5	5	5	0	1	0	1	0
Lookback	3y	1y	5y	3y	5y	1y	3y	1y	3y	1y	1y	1y	1y	1y	1y	1y	1y	1y	1y	1y
Rebalancing	w	w	w	w	w	q	y	y	y	y	q	q	y	q	y	w	m	m	w	w
Buckets	5	5	5	10	10	10	10	10	10	10	5	10	10	5	5	5	10	10	10	10
Weighting	ew	ew	ew	ew	ew	vw	vw	vw	vw	vw	ew	ew	ew	ew	ew	ew	ew	ew	ew	ew

The comparison of winner and loser portfolios provides a more detailed evaluation of each performance metric. Similar to the assessment of the Sharpe and Information ratios in Section 3.4, we can determine which decisions enhance or weaken performance. If a decision variable appears more frequently in winner portfolios and less frequently in loser portfolios (W and L entries are green), it indicates an improvement in performance for that metric. Conversely, if it is more common in loser portfolios and less common in winner portfolios (W and L entries are red), it signals a decline in performance. Additionally, this comparison allows us to assess how decisions impact the dispersion of results. If a variable is overrepresented in the winner and loser portfolios (W entry is green and L entry is red), it leads to greater dispersion in outcomes. In contrast, if a variable is underrepresented in the winner and loser portfolios (W entry is red and the L entry is green), it suggests that the variable leads to a more persistent performance across portfolios. Therefore, beyond identifying performance-enhancing characteristics, this analysis also reveals which decision variables contribute to greater variability or more stable results.

The results of this analysis can be summarized as follows:

- **Risk Estimator:** Portfolios based on idiosyncratic volatility improve mean returns and information ratios, while those based on *beta* result in lower average returns. Interestingly, *beta*-sorted portfolios reduce volatility, whereas idiosyncratic volatility-sorted portfolios increase it. Total volatility estimators (*hist* and *ewma*) perform best in terms of Sharpe ratios, whereas the *beta* estimator performs the worst. This is in line with Liu et al. (2018) who claim that the low-volatility anomaly is stronger using volatility estimates than betas. Additionally, *beta*-sorted portfolios show a wider dispersion in maximum drawdown.
- **Portfolio Type:** With the exception of maximum drawdown and information ratio, where limited-short portfolios result in worse performance, they increase dispersion across all other performance metrics. This further emphasizes the heightened impact of nonstandard errors as the degrees of freedom increase.
- **Price Exclusion:** Price filters generally improve results, increasing the chance of being among winning portfolios and reducing the chance of being among losers across all metrics except the information ratio.
- **Size Exclusion:** Size filters significantly reduce the dispersion of the results for average returns, alpha, and volatility while also improving Sharpe ratios and maximum drawdowns.
- **Lookback Window:** There is no clear winner regarding the length of the lookback window. Generally, shorter lookback windows lead to greater dispersion in outcomes, while longer windows tend to reduce it.
- **Rebalancing Frequency:** Less frequent rebalancing schedules (quarterly and annual) are better for returns, alpha, Sharpe ratio, and information ratio, but worse for reducing volatility. Monthly and quarterly rebalancing perform best in terms of maximum drawdown.

- **Portfolio Buckets:** Larger portfolios (e.g., terciles) are least common among losers and often among winners, indicating generally better performance.
- **Weighting Scheme:** As observed previously, the results for equally-weighted versus value-weighted portfolios are mixed. Equal-weighted portfolios increase the dispersion of results with regard to average returns, alpha, Sharpe ratio, and information ratio. However, for volatility, equally-weighted portfolios are the clear winners, whereas for maximum drawdown, value-weighted portfolios perform best.

The results on portfolio construction choices for winners and losers are robust over time and during recessions. In Appendix C, we confirm this through time-series analyses, which align with the overall findings presented here. Thus, the long-term results are not driven by single events, and top (bottom) portfolios consistently outperform (underperform) over time.

C Robustness Over Time

Up to this point, we have only examined the performance of portfolios over the full time horizon. However, it is crucial to determine whether the top-performing portfolios achieve their success through consistent outperformance or benefit from a one-time event that skews results over the sample period. Furthermore, as low-risk portfolios are designed to prioritize risk reduction, evaluating their performance during periods of financial distress and heightened uncertainty is particularly important.

To go after these two concerns, we perform two analyses that are similar to the analysis in Table B.1. First, to see whether the performance of portfolios is consistent over time, we sort all portfolios based on their realized Sharpe ratios each month for each portfolio type separately and track whether a portfolio is in the upper ("winner") or lower ("loser") quartile of portfolios. Table C.1 holds the relative occurrence of construction choices of the top and bottom 100 winner (W) and loser (L) portfolios. On average, a winning (losing) portfolio occurred 44% (46%) of the time in the upper (lower) quartile of Sharpe ratios, indicating that certain construction choices result in consistent out/underperformance rather than being caused by a one-time event.

Table C.1: Relative appearance of portfolio construction choices in top 100 winner (W) and loser (L) portfolios with transaction costs over time

The table shows the relative occurrence of decision variables in the top 100 winner (W) and bottom 100 loser (L) portfolios, ranked based on consistently high or low Sharpe ratios over time. Portfolios are evaluated net of transaction costs. Each month, portfolios are sorted into quartiles based on their Sharpe ratios. The portfolios that most frequently appear in the top quartile (Q4, "winners") and the bottom quartile (Q1, "losers") over the evaluation period are identified. From these, the 100 portfolios with the highest frequency in Q4 (winners) and Q1 (losers) are selected. On average, the winning portfolios are in Q4 43% of the time, while the losing portfolios are in Q1 48% of the time. For the top and bottom 100 portfolios, the relative occurrence of decision variables is calculated for each decision fork. The analysis is conducted separately for each portfolio type: long-only (*100-0*) and limited-short (*130-30*). Results are highlighted in (light) green if they are more than (two) three standard deviations away from an uniform distribution of variables in a positive direction. For winning portfolios, this indicates that the variable is overrepresented, while for loser portfolios, it suggests the variable is underrepresented. Conversely, we apply (light) red coloring for the opposite scenario. The performance metrics are calculated based on 11,598 daily out-of-sample returns between January 1978 to December 2023.

	100-0		130-30	
Variable	W	L	W	L
Risk Estimator				
hist	6%	0%	1%	3%
ewma	3%	0%	7%	7%
ff3	0%	0%	0%	3%
ff5	0%	0%	0%	4%
beta	91%	100%	92%	83%
Size Exclusion				
0	45%	79%	40%	78%
10	20%	14%	28%	11%
20	35%	7%	32%	11%
Prize Exclusion				
0	28%	41%	29%	49%
1	32%	38%	33%	32%
5	40%	21%	38%	19%
Lookback Window				
1y	21%	53%	46%	57%
3y	38%	27%	26%	31%
5y	41%	20%	28%	12%
Rebalancing Frequency				
w	18%	35%	1%	58%
m	3%	40%	1%	30%
q	28%	15%	41%	9%
y	51%	10%	57%	3%
Portfolio Buckets				
3	13%	10%	51%	6%
5	42%	15%	29%	18%
10	45%	75%	20%	76%
Weighting Scheme				
vw	0%	44%	0%	35%
ew	100%	56%	100%	65%

Comparing the results from this table to the Sharpe ratio column of Table B.1 we find that results are relatively similar. Consistent with the full-horizon findings, size and price filters enhance performance, particularly by mitigating downside losses. Longer lookback windows and rebalancing frequencies improve long-term performance, less extreme portfolio buckets bolster results, and equal weighting leads to more extreme portfolio outcomes.

Contrary to the full-horizon findings, the time-series analysis reveals more extreme outcomes for the choice of the low-risk estimator. Nearly all winning and losing portfolios in the long-only and limited-short portfolios are sorted on *beta*. While *beta*-sorted portfolios account for most losers in the full-horizon analysis in Table B.1, only a few *beta* portfolios emerge as winners. This suggests that *beta*-sorted portfolios exhibit cyclicality, outperforming in certain periods while underperforming in others. This is intuitive when considering *beta*-sorted portfolios as deleveraged market portfolios, which typically underperform during economic upswings and outperform during downturns. In contrast, other risk estimators lack this interpretation and are therefore less likely to exhibit such cyclical behavior.

Second, to assess portfolio performance during periods of market turmoil, we construct Table B.1 for all NBER recession periods in our sample combined. As before, we analyze the top 100 winner (W) and loser (L) portfolios in Table C.2 using various performance metrics. In this analysis, alpha, generalized alpha and maximum drawdown are omitted due to the need for uninterrupted time-series data for these two metrics. Instead, we count the number of times that our low-risk portfolios had a higher return than the benchmark, $r_{p,t} \geq r_{bm,t}$, to capture the consistency of relative outperformance or underperformance of our strategies.

As before, we observe that many construction choices, such as portfolio type, size and price filters, lookback windows, and weighting schemes, yield results during recession periods consistent with the full time-horizon analysis. However, during financial distress, more frequent rebalancing proves beneficial for adapting to market changes. In contrast, the performance of portfolio buckets is less conclusive compared to the clear advantage of using less extreme buckets in the full-horizon analysis. Among estimators, the *ewma* estimator performs better during recessions, which is intuitive given its high respondence to changes in the underlying data. The performance of other risk estimators remains relatively stable.

Overall, the time-series analysis confirms that the favorable construction choices for low-risk portfolios are not tied to one-off events. These choices consistently deliver outperformance over time and provide a valuable edge during periods of financial distress.

Table C.2: Relative appearance of portfolio construction choices in top 100 winner (W) and loser (L) portfolios for long-only and limited-short portfolios with transaction costs during recessions

The table presents the relative occurrence of decision variables in the top 100 winner (W) and bottom 100 loser (L) portfolios, ranked according to various performance metrics during periods of recessions. These portfolios, encompassing both long-only and limited-short strategies, are evaluated net of transaction costs. The relative appearance of decision variables is calculated for each decision fork across the following performance metrics: number of times that the portfolio had a higher return than the benchmark, $r_{p,t} \geq r_{bm,t}$, average return, alpha, volatility, Sharpe ratio, and information ratio. A value-weighted market portfolio serves as the benchmark. Results are highlighted in (light) green if they are more than (two) three standard deviations away from an uniform distribution of variables in a positive direction. For winning portfolios, this indicates that the variable is overrepresented, while for loser portfolios, it suggests the variable is underrepresented. Conversely, we apply (light) red coloring for the opposite scenario. The performance metrics are calculated based on 1,216 daily out-of-sample returns in recession periods between January 1978 to December 2023.

	$r_{p,t} \geq r_{bm,t}$		Return		Volatility		SR		IR	
Variable	W	L	W	L	W	L	W	L	W	L
Risk Estimator										
hist	8%	2%	22%	7%	0%	0%	23%	3%	41%	9%
ewma	27%	0%	62%	14%	8%	0%	62%	13%	4%	14%
ff3	3%	53%	13%	5%	0%	46%	11%	1%	29%	9%
ff5	6%	45%	3%	5%	0%	54%	2%	1%	26%	9%
beta	56%	0%	0%	69%	92%	0%	2%	82%	0%	59%
Portfolio Type										
100-0	74%	39%	0%	18%	7%	35%	0%	17%	93%	21%
130-30	26%	61%	100%	82%	93%	65%	100%	83%	7%	79%
Size Exclusion										
0	16%	60%	40%	95%	90%	87%	37%	93%	18%	97%
10	29%	28%	35%	3%	9%	13%	32%	5%	40%	3%
20	55%	12%	25%	2%	1%	0%	31%	2%	42%	0%
Prize Exclusion										
0	38%	41%	27%	51%	30%	73%	26%	42%	40%	57%
1	34%	32%	33%	34%	37%	27%	35%	38%	43%	33%
5	28%	27%	40%	15%	33%	0%	39%	20%	17%	10%
Lookback Window										
1y	25%	8%	53%	58%	43%	3%	52%	59%	3%	57%
3y	24%	26%	28%	29%	32%	34%	29%	26%	49%	27%
5y	51%	66%	19%	13%	25%	63%	19%	15%	48%	16%
Rebalancing Frequency										
w	18%	45%	0%	62%	40%	26%	0%	52%	67%	69%
m	35%	14%	50%	21%	23%	24%	50%	26%	7%	20%
q	35%	8%	21%	14%	19%	23%	17%	14%	5%	9%
y	12%	33%	29%	3%	18%	27%	33%	8%	21%	2%
Portfolio Buckets										
3	93%	49%	0%	13%	12%	62%	0%	11%	34%	23%
5	4%	43%	6%	29%	39%	19%	5%	28%	62%	29%
10	3%	8%	94%	58%	49%	19%	95%	61%	4%	48%
Weighting Scheme										
vw	40%	82%	92%	0%	0%	1%	87%	0%	8%	0%
ew	60%	18%	8%	100%	100%	99%	13%	100%	92%	100%

D Additional Table

Table D.1: Relative appearance of portfolio construction choices in the rolling ensemble portfolio

The table presents the relative occurrence of decision variables in our rolling ensemble portfolio. The relative appearance of decision variables is calculated for each decision separately. Results are highlighted in (light) green if they are more than (two) three standard deviations away from an uniform distribution of variables in a positive direction. This indicates that the variable is overrepresented. Conversely, we apply (light) red coloring for the opposite scenario.

Variable	Rel. Occ.
Risk Estimator	
hist	27%
ewma	36%
ff3	1%
ff5	1%
beta	35%
Portfolio Type	
100-0	14%
130-30	86%
Size Exclusion	
0	22%
10	46%
20	32%
Prize Exclusion	
0	27%
1	29%
5	44%
Lookback Window	
1y	54%
3y	22%
5y	24%
Rebalancing Frequency	
w	8%
m	3%
q	34%
y	54%
Portfolio Buckets	
3	30%
5	44%
10	26%
Weighting Scheme	
vw	2%
ew	98%