The Mixed vs the Integrated Approach to Style Investing: Much Ado About Nothing?*

Markus Leippold† Roger Rueegg‡

April 21, 2017

ABSTRACT

Abstract

We study the difference between the returns to the integrated approach to style investing and those to the mixed approach. Unlike the mixed approach, the integrated approach aggregates factor characteristics at security level. Recent literature finds that the integrated approach dominates the mixed approach. Using statistical tools for robust performance testing, we demystify these findings as a statistical fluke. We do not find any evidence favoring the integrated approach. What we do find is that the integrated approach exhibits a higher sensitivity to the low-risk anomaly. However, this reduction in risk does not lead to an improvement in performance.

JEL classification: G11, G12, G14

Key Words: Factor investing, integrated and mixed approach, value, momentum, low volatility.

*We thank Michael Wolf for insightful comments and Fabian Ackermann, Michael Bretscher, Andreas Kappler, Florian Arnold, and Andri Silberschmidt for helpful discussions.

†Department of Banking and Finance, University of Zurich

‡University of Zurich and Zurich Cantonal Bank

1 Introduction

Style investing is the investment process that aims to harvest risk premia through exposure to factors. Factors are the foundation of all portfolios: they are the persistent forces driving the returns of stocks, bonds, and other assets. There are diverging views on how to build multi-factor portfolios. The current debate is centered around two approaches. The first approach is to mix where a portfolio is built by combining stand-alone factor portfolios. The second approach is to integrate where a portfolio is built by selecting securities that have simultaneously strong exposure to multiple factors at once. Recent research suggests that a bottom–up or integrated approach provides higher returns and lower risks than a mixed approach.[1] Hence, it seems that the debate about mixing or integrating has been concluded.

However, such a finding clearly must invite suspicion, as it contradicts the standard paradigm in finance. Higher returns can only be achieved by taking higher risks.[2] We contribute to the recent literature on style investing by providing a thorough analysis of the differences in the returns and risk of the mixed and integrated approaches to long-only style investing. We find that the integrated approach shows superior returns to risk characteristics in only a few combinations of styles. When we adjust for multiple hypothesis testing, we can no longer reject the hypothesis that the two approaches are the same. Hence, our findings present a challenge to the previous literature that promotes the integrated approach.

An early contribution that analyzes an integrated (or bottom–up) approach to style investing is Haugen and Baker (1996). With a selected set of factors, they show that factor models are surprisingly accurate in forecasting the future relative returns of stocks. They find high abnormal returns together with lower risk numbers in stocks with high predicted returns and argue that their result reveals a major failure in the efficient markets hypothesis. Subsequent contributions on style investing turned their focus on how to optimally combine individual factor portfolios. Of interest was not whether to

[1] See Bender and Wang (2016), Clarke et al. (2016), and Fitzgibbons et al. (2016).

[2] In addition, it contradicts the risk based explanation of why style premia exist. To clarify this point, consider a combination of value and momentum stocks. Strictly speaking, we thereby avoid overvalued momentum stocks that are threatened by a sudden market crash. However, bearing the risk of a sudden crash is the most rational explanation of why the momentum premium exists (see Daniel and Moskowitz (2015)). Hence, we would expect lower returns in the integrated approach, contrary to what recent publications suggest.

mix or integrate, but on how to derive optimal factor exposures.[3] Factor portfolios were regarded as given group constraints.

Only recently, the integrated approach regained attention with two recent publications. Clarke et al. (2016) argue that the mixed approach captures only one-half of the potential improvement over the market Sharpe ratio. They show that when the group constraint is released and the securities are viewed as a bundle of styles instead of the styles being regarded as a bundle of securities, one can capture much more of the excess returns of the factors. The second work promoting the integrated approach is Bender and Wang (2016). They assert that integration leads to a superior risk–return trade-off due to the fact that it captures nonlinear cross-sectional interaction effects between factors.

Interestingly, the ETF industry has yet to make up its mind whether mixing or integrating is the right approach. Table 1 summarizes the most well-known multi-factor ETFs. While the largest ETF, managed by Goldman Sachs, pursues a mixed approach to factor investing, we find FlexShares, JP Morgan, and iShares implementing an integrated approach to factor investing. Moreover, AQR, one of the largest global investment managers with $159.2 billion assets under management as of August 2016, maintain in Fitzgibbons et al. (2016) that a long-only portfolio is more profitable if based on an integrated approach. Indeed, it is the general tenet of the financial industry that the integrated approach is superior to the mixed approach. To our best knowledge, the only contrarian view that we are aware of is the white paper by Fraser-Jenkins et al. (2016). They find that the integrated and mixed approaches lie on the same return to risk line.

[Table 1 about here.]

Given the inconclusive evidence, we conduct an in-depth analysis of the two long-only methodologies of style investing and contribute to the literature in several ways. First, we analyze all the combinations of the Fama and French (2015) five-factor model extended by the momentum and low volatility factors. Moreover, we analyze an extended period from 1963 to 2016 of all NYSE, AMEX, NASDAQ stocks. By doing so, we expand on the previous literature that concentrates on only a few combinations of styles and markets, mostly on a shorter time period. For example, Fitzgibbons et al. (2016) analyze the combination of value and momentum from 1993 to December 2015, and Bender

[3] See, e.g. Blitz (2015) for an overview.

and Wang (2016) the six possible two- and four-factor combinations of value, low volatility, quality, and momentum, from 1993 to March 2015. Clarke et al. (2016) analyze the four-factor combination of low beta, size, value, and momentum, from 1968 to 2014.

Second, and more importantly, we extend the comparison of the two approaches by building a robust multiple hypothesis framework. While the previous literature reports simple risk and return differences and finds economically sound advantages to the bottom–up construction, we question these results. Motivated by Bailey et al. (2014), who argue that shallow statistical analysis can easily lead to allocating capital to strategies that were false discoveries, we apply a set of robust performance tests to the hypothesis that the integrated approach offers a better performance than the mixed approach. To avoid backtest overfitting, we adjust for the number of portfolio combinations tried. Such an adjustment is common in medical research. Yet, in finance, multiple hypotheses methods have only recently gained attention.[4] Hence, we hope that our study will increase the awareness that the lack of rigorous statistical procedures might lead to wrong and misleading conclusions.

Our empirical results are as follows. When we follow the argumentation of the previous literature, we can confirm their results in that the integrated approach is superior. However, when we apply our battery of more robust statistical tests and include all possible style combinations as well as a longer time horizon, we must conclude that the performance differences are statistically insignificant. What we also find is that the risk of the integrated approach may be lower than that of the mixed approach. Furthermore, it turns out that the integrated approach shows a high sensitivity to the low-risk anomaly, originally discovered by Jensen et al. (1972). This result confirms our intuition behind the integrated approach, which is one of avoiding risk through broader diversification. However, we also find that the lower risk is accompanied by lower returns. Hence, the risk reduction of the integrated approach does not lead to an improvement of performance. When we further analyze trading costs and turnover, we find on average a lower turnover in the integrated approach. This can lead to significant differences in selected portfolio construction techniques and style combinations when trading costs are high. However, due to the low trading costs nowadays we observe no significant difference in the most recent past.

The paper proceeds as follows. In Section 2, we present the methodology behind our portfolio

[4] See, e.g., Harvey et al. (2016) for a current discussion.

construction and hypothesis testing. Section 3 presents our data and factor choice. In Sections 4, we summarize our empirical findings. In Section 5, we provide a turnover analysis and point at possible limitations. Section 6 concludes.

2 Methodology

We now provide some guidance on the methodology for constructing portfolios in the mixed and integrated approach. Then, we briefly present the statistical framework for testing whether the integrated approach to style investing offers higher risk-adjusted excess returns than the mixed approach.

2.1 Portfolio construction

We assume that we have $i = 1, ..., n$ securities and $f = 1, ..., k$ styles with the style information matrix $\Phi \in \mathbb{R}^{n,k}$. Each column $\phi_f \in \mathbb{R}^n$ of Φ contains the style figures for the n securities. For example, for the style 'value' these figures are the book-to-market ratios of the companies. Each security obtains for each factor a score $s_{i,f}$ based on the style information. There are two common ways to build this score, the rank-based and z-score approach. The rank-based score neglects the distribution of ϕ_f and scores the securities among their ranks. We build the score as

$$s_{i,f}^{\text{rank}}(\phi_f) = \frac{\text{rank}(\phi_{f,i}, \phi_f) - 1}{n - 1}, \tag{1}$$

where the operator 'rank' runs from 1 to n from the smallest to the largest values in ϕ_f. The score is invariant to the numbers of securities and lies between 0 (worst) and 1 (best). On the other hand, the distribution of ϕ_f is taken into account in the z-score approach. Here, the score is defined as

$$s_{i,f}^{z} = \frac{\phi_{f,i} - \mu(\phi_f)}{\sigma(\phi_f)}. \tag{2}$$

In the mixed approach, we express the single style portfolios $w_f \in \mathbb{R}^n$ as a function φ of the score vector s_f^j,

$$w_f = \varphi(s_f^j), \tag{3}$$

where $j = \{\text{rank}, z\}$. These portfolio weights are then aggregated to the final weights of the mixed approach by giving a weight of a_f to each style portfolio:

$$w_{\text{mix}} = \sum_{f=1}^{F} a_f w_f. \tag{4}$$

In the integrated approach, the aggregation of the style information occurs before constructing the portfolio. For this purpose, we build an aggregated score as follows:

$$s^j_{\text{agg}} = \sum_{f=1}^{F} a_f s^j_f, \tag{5}$$

where we set the weight a_f of each score equal to the style factor portfolios' weight of the mixed approach.[5] To build the integrated portfolio, the same portfolio construction function φ is applied as for the single style portfolios, but the input score vector is the aggregated score s^j_{agg}:

$$w_{\text{int}} = \varphi(s^j_{\text{agg}}). \tag{6}$$

The main difference between the mixed and integrated portfolios is that the mixed portfolio starts with style portfolios based on single scores and then aggregates the information by mixing the style portfolios. In the integrated approach, the information aggregation occurs before constructing the portfolio. Basically, we are free to choose the portfolio construction function φ and score methodology. While for the score methodology, we restrict ourselves to the rank and z-score, we apply four different portfolios methodologies. The first two are analogous to Fama and French (1992) (TER and DEC) and, for the benchmark-sensitive investors, we additionally include the portfolio construction techniques of Bender and Wang (2016) (BW) and Fitzgibbons et al. (2016) (TE) into our analysis. Hence, we end of with the following set of portfolio construction methodologies,

$$\mathcal{P} = \{\text{TER},\ \text{DEC},\ \text{BW},\ \text{TE}\}, \tag{7}$$

which we briefly discuss next.

[5] Concerning the choice of the style portfolio and score weights a_f, we follow DeMiguel et al. (2009) and apply the most naive diversification rule $a_f = \frac{1}{F}$.

2.1.1 Tercile and decile portfolios: TER and DEC

The tercile (TER) and decile (DEC) portfolio construction follows closely the style portfolio construction originally suggested by Fama and French (1992). First, we build the scores with the rank-based methodology in Equation (1). Second, the function φ of Equations (4) and (6) is such that we invest value-weighted in the upper tercile of the scored companies for the TER and in the upper deciles of the scored companies for the DEC approach.

To clarify the differences between the mixed and integrated approaches, we provide a stylized example for the TER portfolios. We assume that there are 10 stocks, from stock A to stock J, with a given market capitalization (mc) and two factors $f = \{V, W\}$, say, 'value' (V) and 'momentum' (M). Exact numbers are shown in Table 2. The three highest ϕ_V and ϕ_M figures are highlighted in bold. The three highest book-to-market ratios are 0.51 for stock A, 0.82 for stock D, and 0.97 for stock I. The highest returns for the past 12 months disregarding the most recent month are 0.09 for stock A, 0.14 for stock B, and 0.22 for stock I. We first build the 'value' and 'momentum' scores s_V^{rank} and s_M^{rank} as illustrated in Equation (1) and take their the average to arrive at the aggregated score $s_{\text{agg}}^{\text{rank}}$ shown in Equation (5).

[Table 2 about here.]

For the mixed portfolio, we first apply the portfolio construction function φ to the scores s_f^{rank} and value-weight the upper tercile of the stocks. This procedure results in the single style portfolios w_V and w_W. The final weights are simply the average of the factor portfolio weights and are shown in column w_{mix}. In contrast, the integrated approach aggregates the information on security level. We first build the aggregated score $s_{\text{agg}}^{\text{rank}}$ that is the average of the style scores s_V^{rank} and s_M^{rank}. Given the aggregated score, we then value-weight the upper tercile of the stocks as shown in column w_{int}.

We end up with four groups of stocks. The first group is assigned a weight of zero in either approach. They show neither superior style characteristics, nor are they on average superior to the other stocks. The second group of stocks is only represented in the mixed portfolio, but not in the integrated approach. They show superior style characteristics in one of the two styles. However, the score in the second style is too small to be considered in the integrated portfolio. The third group of

stocks is not considered in any of the style portfolios, but it shows superior style characteristics when all styles are aggregated. Stock C is an example of this group. The fourth group of stocks belongs in both the mixed and integrated portfolios.

The mixed portfolio always holds at least as many stocks as the integrated portfolio. The more similar the styles, the fewer stocks are included in the mixed portfolio. If each style provides the exact same ranking of stocks, the weights of the integrated and mixed approaches are equal. With two (three) styles and without an overlap, it is possible to hold 60 (90) percent of the stocks in the mixed portfolio. For more than four styles, the mixed portfolio could possibly hold all the stocks of the universe, while on the other hand the integrated approach holds by definition in any case 30 percent of the stocks in the universe.

2.1.2 Bender and Wang (2016) portfolios: BW

In contrast to the tercile and decile portfolios, the portfolio construction of Bender and Wang (2016) concentrates on under- and overweighting relative to the market-cap-weighted benchmark. The scores are built on the z-score methodology in Equation (2). The portfolio construction function φ is defined by first ordering the securities according to the score and grouping them into 20 subportfolio with each holding 5 percent of the total market capitalization. For each group, a multiplier of 0.05 to 1.95 with increments of 0.1 is applied to the market-cap weight of the securities. For example, the subportfolio of companies with the highest (lowest) score gets its market-cap weight multiplied by 1.95 (0.05). In the last step, the weights are normalized. This procedure results in an over- and underweighting of the highest and lowest scored companies. For a detailed description of the resulting portfolios, we refer to the original work of Bender and Wang (2016). Compared to the TEC and DEC portfolios of the previous section, the BW portfolios invest in all securities in the mixed and integrated approach.

2.1.3 Target tracking error portfolios: TE

We additionally implement the target tracking error optimization suggested in Fitzgibbons et al. (2016). The portfolio construction function φ is defined as

$$\varphi(s) := \max_w (w - w_M)'s \quad \text{s. t.} \quad \sqrt{(w - w_M)'\Sigma(w - w_M)} \leq \sigma_{te}, \tag{8}$$

where we construct the score s with the z-score methodology in Equation (2) for the single style portfolios of the mixed approach and with the aggregated score in Equation (5) for the integrated approach. Furthermore, by σ_{te} we denote the ex-ante target tracking error, by w_M the market-cap-weighted benchmark, and by Σ the covariance matrix of the returns. Since we deal with a large covariance matrix, we use the shrinkage methodology of Ledoit and Wolf (2004) based on the most recent 24 observations. To obtain similar levels of tracking errors as in the BW approach, we set the ex-ante target tracking error σ_{te} to two percent annualized.

2.2 Multiple hypothesis testing

As Bailey et al. (2014) argue, researchers and financial institutions are incentivized to try several possibilities, but report only the significant results. In our case, we have 5 styles that result in 26 possible combinations. Therefore, our hypothesis is split into 26 individual hypotheses. To make the case for multiple hypothesis testing, we provide a simple illustration with a momentum-based strategy. To this end, we consider IBM stock with a history of returns data from January 1960 to December 2014. We create 20 momentum strategies. The first strategy invests in IBM for the next month if the previous month was positive, otherwise it steps out of the market. The second strategy analogously invests in IBM for the next month if the second most recent month was positive, and disinvests if it was negative. This analysis is conducted for the most recent 20 months, which results in 20 different momentum strategies. Our main goal is to test whether a specific momentum strategy shows a significantly higher Sharpe ratio than the buy-and-hold strategy. For the test statistics, we consider the robust Sharpe ratio test of Ledoit and Wolf (2008).

When testing 20 different momentum strategies as specified above, we have to be sensitive to

the fact that the probability of finding no significant results at a confidence level of 5 percent is $(1-0.1)^{20}$ or approximately 12 percent. Therefore, the complementary probability of finding at least one significant momentum strategy by pure luck is roughly 88 percent. By inspection of Table 3, we find that there are three Sharpe ratios significantly different from the buy-and-hold strategy, two significantly lower and one significantly higher. In particular, the return of the momentum strategy with the 16th look-back month shows an annualized return of 7.1 percent with a annual volatility of 16.5 percent. For comparison, the annualized return of IBM over the period June 1963 to December 2014 was lower, at 4 percent, with a higher annualized volatility of 23.8 percent. We can reject the hypothesis that the Sharpe ratio is equal to the buy-and-hold strategy at a confidence level of 10 percent. However, recalling that we tested 20 different strategies and to be sure that our superior strategy is not just a statical fluke, we must embed our p-values into a robust multiple hypothesis framework.

[Table 3 about here.]

There are many ways to deal with the problem of multiple hypothesis testing.[6] For our analysis, we focus on the false discovery rate (FDR). The FDR was introduced in Benjamini and Hochberg (1995) and is defined as the expectation of the proportion of falsely rejected null hypotheses. But Romano and Wolf (2005a) and Romano and Wolf (2005b) point out that they make the strong assumption that the individual p-values are independent of each other. Therefore, they propose a resampling-based stepdown multiple testing framework that considers the dependence structure of the test statistics. Their method comes at a high computational cost. In Romano and Wolf (2016), the authors refine their method. Not only are they able to reduce the computational cost, but their method also avoids choosing a fixed significance level α. Hence, their framework allows for dependence structures in the test statistics without loss in statistical power.

For comparison and as a robustness test, we also focus on the older tests of Bonferroni (1936) and Holm (1979) that control the familywise error rate (FWER), the probability of at least one false discovery. The Bonferroni test divides the required significance value by the number of hypotheses. A confidence level of 5 percent with 20 tries produces a threshold of $0.05/20 = 0.0025$. As pointed out

[6] For a comprehensive overview, we refer to Harvey et al. (2016).

by Conneely and Boehnke (2007), the Bonferroni adjustment is too conservative for correlated tests. Holm (1979) developed a sequential Bonferroni method, preserving its flexibility but increasing its power. Strictly speaking, the test applies the Bonferroni adjustment only on the subset of hypotheses that are not rejected from the beginning.[7]

How does our previous conclusion from the analysis of the momentum strategy change if we adjust the single hypothesis p-values for multiple hypothesis testing? From Table 3, we observe three significant p-values when we test naively; however, with the multiple hypothesis adjustment, the p-values are all insignificant. Therefore, we can not reject the hypothesis that one of the momentum strategies is different from a simple buy-and-hold strategy. Hence, their abnormal performance is probably a false discovery and nothing but a statistical fluke.

3 Data

We now present our data and factor choices for the multiple hypothesis testing.

3.1 US data from 1963 to 2016

To construct the integrated and mixed portfolios, we use stock return and balance sheet data from the merged CRSP and Compustat database. The stock universe consists of all NYSE, AMEX, and NASDAQ stocks with share codes 10 or 11. The data items of the CRSP and Compustat database are merged by the eight-character CUSIP. We exclude finance, insurance, and real estate companies with SIC codes between 6000 and 6799. We limit the universe to big stocks as defined by Fama and French (1992). This universe consists of 810 stocks on average, starting with the 583 largest US companies in June 1963 and ending with the largest 867 companies in December 2016. At its peak in February 2000, there were 1,548 companies in the universe. The market capitalization breakpoint, which splits the universe into small and large caps, builds the median of all NYSE stocks. The average market capitalization over the analyzed period was 792 million. It reached a minimum of 63 million in December 1974 and peaked in June 2014 with 2,871 million. The universe represents the

[7] Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) show that under dependency, it is favorable to control for the FDR, and not for the FWER.

largest companies in the most important equity market of that time. Consequently, it is a highly liquid universe, tradeable with small transaction costs.

The CRSP database has a monthly frequency and starts in January 1960. As outlined by Fama and French (1992), the data of the Compustat database is not reliable before 1962. Therefore, we use balance sheet information at a yearly frequency from the Compustat database starting at the earliest possible year, in 1962. Moreover, all data points of the Compustat database are lagged by 6 months to guarantee that the balance sheet data of the companies are available at the date of the portfolio construction. Considering the lag of 6 months for the balanced sheet data, the backtest period consists of 63.5 years. It starts in June 1963 and ends in December 2016.

3.2 Factor choices and factor combinations

The most prominent factor model is Fama and French (1992) with the three factors market excess return (M), 'size' (S), and 'value' (V). Carhart (1997) extended the Fama and French factors with the 'momentum' factor (W) of Jegadeesh and Titman (1993). The recent literature has put a lot of effort into detecting other factors that show high abnormal returns. Harvey et al. (2016) find 315 published factors with ostensibly significant excess returns. Among the most recent findings are the 'quality' premium as defined by profitability, growth, safety, or payout prevail anomalies.[8] Consequently, Fama and French (2015) extended their three-factor model to include certain quality aspects with the 'profitability' factor of Novy-Marx (2013) and the 'investment' factor of Aharoni et al. (2013). They argue that expected returns are not solely driven by the book to market ratio (V), but also by 'profitability' (R), and 'investment' (C). Another highly popular anomaly is the 'low-risk' (L) anomaly, originally discovered by Jensen et al. (1972). Empirically, low-beta stocks exhibit higher returns than implied by their market beta. Among many other low-risk measures, Ang et al. (2006) find that stocks with high idiosyncratic risk earn abnormally low average returns.[9]

Without loss of generality, we focus on the most important and widespread style factors: 'value', 'profitability', 'investment', 'momentum', and 'low volatility'. We argue that if the integrated ap-

[8] A good overview of the quality factor is given by Asness et al. (2014).

[9] Recent overviews on the low-risk anomaly include, e.g., Blitz and Van Vliet (2007) and Baker et al. (2011). As a result of the high risk-adjusted returns of the low-risk stocks, every index or smart-beta provider offers a low-volatility product. Data compiled by Bloomberg show that the 10 largest low-volatility or minimum volatility ETFs held $40 billion in assets as of mid-2016.

proach obtains higher risk-adjusted returns compared to the mixed approach, higher risk-adjusted returns should also be observed for the combinations of these most popular factors. Moreover, we concentrate on independent risk factors to arrive at a meaningful analysis. Hence, our analysis is built on the following set of factors $\mathcal{F}$:

$$\mathcal{F} = \{V, R, C, W, L\}. \tag{9}$$

We measure V by book equity as defined in Fama and French (1992) divided by the market capitalization of the CRSP database. We lag book equity by 6 months in order to guarantee that the balance sheet data is published at the date of the portfolio construction. In contrast, the market capitalization is not lagged. Asness and Frazzini (2013) show that this small detail is superior in terms of performance and when the portfolio is rebalanced monthly. The factors R, C, and W are defined as in Fama and French (2015). R is calculated by the ratio of operating profitability divided by book equity and C by the total book assets of the recent year divided by the actual total book assets. All these variables are calculated with data from the Compustat database and lagged by 6 months. W is calculated as the total returns over the past 12 months, while the most recent month is ignored. For L we take the volatility to be the standard deviation of the most recent 36 monthly returns. The look-back period of 36 months is chosen analogously to Blitz and Van Vliet (2007). We show the summary statistics of the factors including the size factor in Table 4.

[Table 4 about here.]

It would be beyond the scope of this paper to test all possible available factors. Instead, we want to give a comprehensive overview of the most popular factors. Moreover the factors should also be independent of each other. For example, Fama and French (1992) also test the earnings-to-price ratio. Since this ratio is highly correlated with the book-to-price ratio, we do not include it in our study, to avoid potential problems arising from multicollinearity. Therefore, before we proceed, we test our selected factors for multicollinearity by calculating the variance inflation factor (VIF). Table 5 reports the results. We find that the VIF stays below two, far below the threshold of 10 which, according to O'Brien (2007), is equivalent to a confidence level of 0.1. Therefore, there is no sign of a linear

dependency in our factor selection (9).

[Table 5 about here.]

Given the 5 following styles, 'value' (V), 'profitability' (R), 'investment' (C), 'momentum' (W), and 'low-volatility' (L), it is possible to build 10 combinations with 2 factors, 10 combinations with 3 factors, 5 combinations with 4 factors, and 1 combination with 5 factors. In all, we end up with 26 possible combinations, denoted by $\mathcal{C}$,

$$\begin{aligned}\mathcal{C} = \{ & VW, VC, VR, VL, WC, WR, WL, CR, CL, RL, \\ & VWC, VWR, VWL, VCR, VCL, VRL, WCR, WCL, WRL, CRL, \\ & VWCR, VWCL, VWRL, VCRL, WCRL, VWCRL\},\end{aligned} \tag{10}$$

where we indicate each combination by the acronym formed from its factors' names.[10]

4 Robust hypothesis testing

We now run the integrated and mixed portfolios for all 26 combinations of $\mathcal{C}$ and all four portfolio construction methodologies in $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We first compare the Sharpe ratio of the resulting 104 strategies and then concentrate on the variance and the benchmark orientated figures, namely the information ratio and the tracking error.

4.1 Multiple hypothesis testing for the Sharpe ratio

In Figure 1, we report the Sharpe ratios of all 104 strategies in grey and highlight the differences of the integrated to the mixed approach in green (positive) or red (negative). The integrated approach obtains a higher Sharpe ratios in most of the cases. For the BW portfolio construction methodology, we find the integrated approach to outperform the mixed approach in any of the style combinations.

[Figure 1 about here.]

[10] For example, the combination of 'value' (V), 'momentum' (M) and 'low-volatility' (L), is called VWL.

Bender and Wang (2016) find the highest difference in risk-adjusted returns for the combination of 'value', 'low volatility', 'quality', and 'momentum', with 0.84 in the integrated approach and 0.73 in the mixed approach. We too find the highest Sharpe ratio for the 4-factor combinations. For example $VWCR$ or $VWRL$ show, from June 1963 to December 2016, very high Sharpe ratios: 0.48 and 0.49. In contrast, the mixed approach obtains a Sharpe ratio of 0.41 for both combinations. When we regard the style 'robust' as a proxy for 'quality', we arrive for the combination $VWRL$ to the same magnitude of improvement as Bender and Wang (2016). Fitzgibbons et al. (2016) find increasing benefits with the number of uncorrelated factors combined in the portfolio's construction for the post-1993 period. We can also find support for this observation. We find only positive differences for combinations with more than three styles, while we observe only small improvements for the two-factor combinations.

Next, we perform the single hypothesis robust Sharpe ratio test of Ledoit and Wolf (2008). The robust Sharpe ratio test requires an optimal block size for the dependent block bootstrap.[11] Since we observe the highest autocorrelation for the block sizes five, we use this value for the optimal block size (bl).[12] Figure 2 shows the monthly Sharpe ratio differences of the four portfolio construction methodologies and the 26 factor combinations in bars as well as their p-values in symbols. The significant differences at the 95 percent confidence level are highlighted in green (positive) or red (negative). For the TER (DEC) construction, we observe only one (three) significant single hypothesis tests. For the benchmark-orientated portfolio construction method BW (TE), we find 16 (15) p-values to be below the five percent level. Hence, at least under a single-hypothesis test, there seems to be some pattern emerging in favor of the integrated approach when applying the benchmark-oriented construction methods.

[Figure 2 about here.]

However, since we deal with four portfolio construction methods and 26 factor combinations, it is crucial to adjust the single hypothesis p-values in Figure 2 for the numbers of tries. For example, Clarke et al. (2016) focus on the ex-ante portfolio construction and the comparison of the two ap-

[11] There are methods to evaluate the optimal block length for a single time series, see Politis and White (2004) and Patton et al. (2009), or in the bivariate case, see Ledoit and Wolf (2008). But for the multivariate case with more than two time series, there is no framework available.

[12] For robustness we also applied a block size of two, that shows the second highest autocorrelation. The different block size has no impact in all the results presented below and lead to the same conclusions.

proaches. They make the – in their words – 'implausible' assumption that the investor first has the knowledge of the successful 4 styles as well as their information ratio. We argue that exactly this implausible assumption is the game changer. First, the expected information ratios are unknown ex ante. It may well happen that the successful styles of the past will fail in the future. Second, the investor is not aware which styles or which combination will be successful in the future. Therefore, the multiple hypothesis framework that we apply next is of crucial importance. We expect the different portfolio construction methods for the same style combination to behave similarly. Since the framework of Romano and Wolf (2016) accounts for these inherent dependence structures in the test statistics, it is best suited to adjust the single hypothesis p-values.

[Figure 3 about here.]

Figure 3 provides the adjusted p-values (in symbols) under the multiple hypothesis testing framework. We find that all the significant strategies of the TER and DEC portfolio construction are likely to be false discoveries. However, for the benchmark-sensitive investors, we still find five combinations including VW in the framework of BW and five combinations including WL in the TE framework to survive the adjustment and to perform significantly better.

4.2 Variance, information ratio, and tracking error comparison

We now test the null hypothesis that the variance, the information ratio, and the tracking error of the integrated and mixed approach are equal for the 26 factor combinations and four portfolio construction methodologies. Figure 4 reports the differences in the logarithmic variance (top chart), the information ratios (middle chart), and the logarithmic tracking error (bottom chart) based on monthly returns using the integrated and mixed approach. For the variance and tracking error hypothesis testing, we apply the robust variance test of Ledoit and Wolf (2011). Analogous to the previous section, we chose the block size of five for the dependent block bootstrap that is also required for the robust variance test. The only difference to the robust hypothesis testing with the Sharpe ratio is that the analyzed returns are the excess returns above the one-month Treasury bill rate (for the information ratio) and the excess returns above the market-cap weighted benchmark (for the variance and tracking error).

[Figure 4 about here.]

We find eleven combinations for the TER, three for the DEC, 15 for the BW, and one for the TE approach that exhibit a lower variance in the integrated approach, while only the combination *VR* shows a significantly higher variance in the DEC portfolio construction methodology. It is not surprising that the optimized portfolio construction method TE results in similar risk levels, since including the information of the covariance matrix during the portfolio construction leads to neutralized bets in the risk dimension. The more factors included in the strategies, the higher is the percentage of significant differences where we can reject the null hypothesis.

For the portfolio construction methodologies TER and DEC, we observe in general a lower information ratio compared to the mixed approach. The combination *VWL* in the DEC and *VWCL*, *VWRL* as well as *VWCRL* in the TER are significantly lower compared to the mixed approach. On the other hand, we observe in the benchmark-orientated approaches BW and TE on average an improvement in the information ratio. However, none is significantly different from zero. Again, when we take 'robust' as a proxy for quality, we can confirm the result of Bender and Wang (2016) that the integrated approach shows a higher information ratio over time. However, these differences not significantly different from zero, with adjusted p-values close to one.

For the tracking error, we first observe that the integrated approach obtains a significantly higher tracking error in all of the combinations. Second, the differences increase with the numbers of factors. This result comes not as a surprise, since we construct the single style portfolios of the mixed portfolio in the same way as the final portfolio of the integrated approach. Consequently, the equal-weighted average of the single style portfolios in the mixed approach shows a significant difference in the active risk of the two portfolios. This finding is due to the high diversification effect of the single style portfolios that exhibit a low correlation among each other.

4.3 Similar active risk: a fair comparison

So far, we do find some support, although weak, for the results presented in recent literature, in that the integrated approach shows a significantly higher Sharpe ratio for some of the style combinations. However, we also find significantly higher tracking errors compared to the mixed approach and no

improvement in the information ratio. Hence, a fair comparison between the integrated and mixed approach demands some further investigation. To this end, we construct the portfolios such that the same level of active risk is achieved in both methodologies.[13]

4.3.1 Portfolio construction

To increase active risk in the mixed approach, we change the portfolio construction function from Equation (3). Instead of investing in the 30 percent best stocks for the single mixed TER style portfolios, we invest market-cap-weighted in the 20 (two factors), 12.25 (three factors), 10 (four factors) and 8.75 (five factors) percent best stocks. We expect that this increased concentration results in higher tracking errors for the mixed portfolios. The DEC portfolio construction approach is neglected, since it already has a high level of concentration in the integrated approach that is hard to generate in the mixed approach. For the BW portfolio construction method, we take the multiplier ranging from 0.05 to 1.95 to the power of two (two factors), three (three factors), five (four factors), and eight (five factors) for the single style portfolios w_f. By doing so, we give a higher overweight to the five percent market-cap groups with a high score and increase the underweight of stocks with a small score, relative to the market-cap-weighted benchmark. Analogous to Fitzgibbons et al. (2016), we increase the ex-ante annual tracking error target to 3.0 (two factors), 3.5 (three factors), 3.8 (four factors), and 4.0 (five factors) in the TE mixed portfolios. The portfolio construction of the integrated portfolio remains the same.

4.3.2 Active risk comparison

We now test the hypothesis of equal tracking errors for the 26 factor combinations and three portfolio construction methodologies presented in the previous section. We show the differences in the logarithmic tracking error and the multiple hypothesis adjusted p-values in Figure 5.

[Figure 5 about here.]

In contrast to Figure 4, there are not only positive but also negative (significant) differences between integrated and mixed approach. Hence, on average, both approaches now have similar active

[13] We thank the referee for pointing us into this direction.

risk and we are able to conduct a fair reward to risk analysis comparison in a next step.

4.3.3 Hypothesis testing

Adjusting for multiple hypothesis, we test the null hypothesis that the Sharpe ratio, variance, or information ratios are equal for the 26 style combinations and the three portfolio construction methodologies TER, BW, and TE from the set $\mathcal{P}$. Results are summarized in Figure 6.

[Figure 6 about here.]

For the Sharpe ratio in the top chart, we find no significant difference for any of the 78 combinations tested. We also find that many of the differences decrease and turn to negative numbers. For example, the BW approach shows only in three out of 26 combinations an improvement in the Sharpe ratio, while we found five positive significant Sharpe ratios in our first try in Section 4.1.

For the variance in the middle chart we find that 22 in the TER, four in the BW and eleven style combinations in the TE portfolio construction method show a significantly lower variance over time. In only seven of the 78 tested combinations, we observe a higher variance over time.

For the information ratio in the bottom chart we find for the TER methodology four out of 26 combinations with a higher information ratio in the integrated approach. In the BW methodology we find the combination *VL* to offer a minor improvement in the information ratio, while the other 25 combinations show lower information ratios. On the other hand, for the TE approach we see 19 of the 26 combinations to offer a higher information ratio. We can reject the null hypothesis that the two approaches are equal only for the five factor combination *VWCRL* of the BW portfolios. This combination shows a significant lower information ratio from June 1963 to December 2016.

We conclude that the significant improvements in the Sharpe ratio from Section 4 were due to the different level in active risk of the integrated to the mixed approach in long-only style investing. When we adjust the mixed approach such that it exhibits the same level in active risk, we can no longer reject the null hypothesis that the Sharpe ratio or information ratio is higher in the integrated approach. We even find one negative significant difference in the information ratio.

4.4 Asset pricing tests

To gain some further intuition about the differences of the mixed and integrated approach, we ask whether the return differences can be explained by the risk factors themselves. We therefore run for every combination in (10) and for the three portfolio construction methodologies TER, BW, and TE in $\mathcal{P}$ the following regression:

$$r_{int,t} - r_{mix,t} = \alpha + \beta_M M_t + \beta_S S_t + \beta_V V_t + \beta_R R_t + \beta_C C_t + \beta_W W_t + \beta_L L_t + \epsilon_t, \tag{11}$$

where $r_{int,t} - r_{mix,t}$ corresponds to the difference in monthly returns between the integrated and mixed approach, M is the market return, S is the small minus big factor of Fama and French (1992), and V, R, C, W, and L are the return differences of the upper tercile compared to the lower tercile within our equity universe defined in Section 3.

[Figure 7 about here.]

The parameter estimations and t-values are presented in Figure 7.[14] For the alpha coefficient, we cannot observe a consistent pattern in the t-statistics. To provide an explanation, we recall our stylized example in Section 2.1.1. On the one hand, the integrated approach increases the exposure to factor returns by penalizing negative characteristics. This property follows from aggregating the characteristics on security level, which keeps us from buying stocks with highly negative characteristics in one of the factors. For example, stocks B and D are not included in the integrated portfolio, because they include a negative factor score in one of the styles. On the other hand, the integrated approach decreases the sensitivity to factor returns by avoiding stocks that have diverging style exposures. For instance, stock B, which is highly sensitive to the momentum value factor, and stock D, which is highly sensitive to the value factor, are not included in the integrated approach. But stock C, with less pronounced style characteristics, is included. The insignificant alpha coefficients provide evidence that these two effects neutralize each other.

[14] We use HC3 t-values. The HC3 is a version of the significance tests based on a heteroscedasticity consistent covariance matrix (HCCM), which are consistent even in the presence of heteroscedasticity of an unknown form. We highlight t-values above 1.96 in green and below -1.96 in red. Moreover, we cap t-values above and below five to achieve a better overview.

When analyzing the different factor sensitivities, we observe that the sensitivity to the market factor M is low and mostly negative for combinations with three and more factors. The S factor, which can be seen as a proxy for illiquidity, is mostly negative with high negative t-values. This implies that the integrated approach loads less on liquidity risk, which may serve as an explanation of the lower expected returns in the long run of the integrated approach. For the factors V, R, C, and W, we find no clear pattern for the sensitivities. In contrast, for L we find a consistent positive sensitivity on the part of combinations with more than three factors, and large t-values. This is in line with the findings in Jivraj et al. (2016), who find high sensitivities to the low volatility factor when comparing the integrated and mixed approach for the style combination of value, momentum, low volatility, and quality. The high sensitivity to the low volatility factor L and the low realized risks of the low volatility factor over the analyzed time period are an explanation for the generally lower risk numbers of the integrated approach. The R-squared of the regressions increases with the number of factors considered, and obtain very high levels with an average R-squared of 0.30 for all style combination and portfolio construction techniques.

5 Turnover analysis

There are some limitations to our analysis. The first concerns the portfolio construction process. There are many ways of constructing a factor portfolio. We have focused on the most natural choices, the tercile portfolios being weighted by market capitalization as well as the benchmark-orientated approaches of Bender and Wang (2016) and Fitzgibbons et al. (2016). These constructions are close to the equilibrium portfolio of the CAPM and thereby minimize illiquidity issues. The second concern relates to potential selection biases. We tried to reduce such a bias by analyzing different factors and combinations as well. Yet, the set of factors and their definition was not know at the beginning of our analysis in 1963. Since the selection bias increases with the number of factors considered, results that depend on a large number of factors must be taken with caution.

A third concern, which is highly relevant from a practical viewpoint, is the turnover of the strategies. Figure 8 illustrates the turnover of the three portfolio construction methods with similar active risk for both the integrated and mixed approach with netting (mix - net) and without netting (mix

- gross). The mixed approach with netting views the turnover as if the single style portfolios are managed in one single mixed portfolio. For the mixed approach without netting, we calculate the turnover as if the single style portfolios are managed individually.

[Figure 8 about here.]

We observe that the integrated approach offers in general a lower (green) turnover compared to the mixed approach. When we compare the turnover reduction in the mixed approach with and without netting, we see that the effect is higher in the benchmark-orientated approaches and smaller in the decile approach. Finally for the style combinations including 'momentum' (W), the turnover in any of the portfolio constructions is substantially increased, while the combination of the quality type style factors 'robust' (R), 'investment' (C), and 'low volatility' (L) show much lower turnover over time.

We now test whether turnover, and therefore trading costs, have an impact on our previous results. We estimate transaction costs by starting with a one percent one-way transaction cost from 1963 to 1975. After the May Day in 1975 and the deregulation of the commission fees,[15] we decrease the transaction costs from 1976 to 2016 at an exponential decay with a mean lifetime of twelve years. This results in similar cost levels as used in different studies, such as, e.g., Keim and Madhavan (1998) and Jones (2002). Moreover, in 2016 the resulting transaction costs are 0.033 percent, which corresponds to the bid-ask spreads of US index funds at that time.[16]

[Figure 9 about here.]

Figure 9 shows the difference in the monthly Sharpe (top) and information ratio (bottom) together with the adjusted p-values of the integrated and mixed approach with netting. Due to the higher turnover of the mixed approach, we observe significant Sharpe ratio (information ratio) differences in the combinations WL, VWL, and $VWCL$ (WL and VWL) in the TE portfolios. Also, the reward to risk figures in the TER and BW increase. However, they are still negative or show high adjusted

[15] On May 1, 1975, brokerages were allowed to charge varying commission rates. Prior to this change, all brokerages charged the same price for stock trades.

[16] The bid-ask spread of index funds corresponds to the expected transaction costs in order to protect current investors from new subscriptions or redemptions.

p-values. Due to deregulation and higher volumes, commission fees and slippage decrease steadily over time. Therefore, we are also interested in the impact of trading costs for the more recent period.

[Figure 10 about here.]

Figure 10 shows the same tests for the period June 1993 to December 2016, with trading costs starting at 0.23 percent.[17] Strikingly, we find that for the period after June 1993, there is no evidence to reject the null hypothesis that the two approaches have the same return to reward ratio. The same conclusion holds for the mixed approach without netting. Hence, our analysis of transaction costs shows, due to the lower turnover, that the integrated approach may well be the better choice, if transaction costs are high. However, with the substantial decrease of these costs over the last decades, this advantage has eroded.

6 Conclusion

We rigorously study the difference in returns between the mixed and integrated approaches to long-only style investing. In the US stock market from 1963 to 2016, we analyze the 26 possible combinations of five styles: value, robustness, investment, momentum, and low volatility under the three portfolio construction methodologies of Fama and French (1992), Bender and Wang (2016) and a target tracking error optimization suggested in Fitzgibbons et al. (2016). While the previous literature concentrates on simple performance comparisons for arbitrary factor combinations and portfolio constructions, we apply a robust statistical testing framework. In contradiction to recent findings and the general tenor in the finance industry, we cannot support the hypothesis that the integrated approach leads to superior reward to risk ratios for any of the 26 tested factor combinations and portfolio construction methods.

We further find evidence that the integrated approach shows lower variances over time. In contrast to previous literature that mostly concentrates on a shorter and more recent time horizon, we find that the lower risk is, on average, associated with lower returns. For the integrated approach, we find a high sensitivity to the low volatility anomaly. By aggregating style information at security level,

[17] This breakpoint is also of interest due to the publication of the Fama-French three-factor model at this time and the studies of Bender and Wang (2016) and Fitzgibbons et al. (2016), which analyze data from 1993 onwards.

the integrated approach reduces risks and avoids extreme stocks that exhibit a high sensitivity to only a few (or only one) styles.

Our results confirm that, when naively tested, some factor combinations show superior return to risk ratios over specific periods. But when we apply a multiple hypothesis framework, we must conclude that none of the differences are significant. This conclusion also holds when we adjust for transaction costs. Given the increasing computational power for conducting multiple backtests and given the fact that financial institutions have incentives to deliver extraordinary results, it is crucial to apply the most advanced statistical testing frameworks. Ignoring the available tools can lead to hasty conclusions and mis-allocation of capital to investment strategies that are false discoveries.

References

Aharoni, Gil, Bruce Grundy, and Qi Zeng, 2013, Stock returns and the miller modigliani valuation formula: Revisiting the fama french analysis, *Journal of Financial Economics* 110, 347–357.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.

Asness, Clifford S., and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Asness, Clifford S., Andrea Frazzini, and Lasse H. Pedersen, 2014, Quality minus junk, *Available at SSRN 2312432* .

Bailey, David H., Jonathan M. Borwein, Marcos L. de Prado, and Qiji J. Zhu, 2014, Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance, *Notices of the AMS* 61, 458–471.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler, 2011, Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly, *Financial Analysts Journal* 67, 40–54.

Bender, Jennifer, and Taie Wang, 2016, Can the whole be more than the sum of the parts? Bottom-up versus top-down multifactor portfolio construction, *Journal of Portfolio Management* 42, 39–50.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of statistics* 1165–1188.

Blitz, David, 2015, Factor investing revisited, *Journal of Index Investing* 6, 7–17.

Blitz, David, and Pim Van Vliet, 2007, The volatility effect: Lower risk without lower return, *Journal of Portfolio Management* 34, 102–113.

Bonferroni, Carlo E., 1936, *Teoria statistica delle classi e calcolo delle probabilita* (Libreria internazionale Seeber).

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

Clarke, Roger G., Harindra De Silva, and Steven Thorley, 2016, Fundamentals of efficient factor investing, *Financial Analysts Journal* 72, 9–26.

Conneely, Karen N., and Michael Boehnke, 2007, So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests, *American Journal of Human Genetics* 81, 1158–1168.

Daniel, Kent D., and Tobias J. Moskowitz, 2015, Momentum crashes, *Journal of Financial Economics* forthcoming.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009, Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?, *Review of Financial Studies* 22, 1915–1953.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fitzgibbons, Shaun, Jacques Friedman, Lukasz Pomorski, and Laura Serban, 2016, Long-only style investing: Don't just mix, integrate, Integrate (June 29, 2016), AQR Capital Management, LLC.

Fraser-Jenkins, Inigo, Alix Guerrini, Alla Harmsworth, Mark Diver, Sarah McCarthy, Robertas Stancikas, and Maureen Hughes, 2016, Global quantitative strategy: How to combine factors? It depends why you are doing it, Global Quantitative Strategy (September 14, 2016), Sanford Bernstein.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ...and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Haugen, Robert A., and Nardin L. Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

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

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

Jensen, Michael C., Fischer Black, and Myron S. Scholes, 1972, *The capital asset pricing model: Some empirical tests* (Praeger Publishers Inc).

Jivraj, Farouk, David Haefliger, Zein Khan, and Benedict Redmond, 2016, Equity multi-factor approaches: Sum of factors vs. multi-factor ranking, QIS Insights (September 16, 2016), Barclays Investment Bank.

Jones, Charles M., 2002, A century of stock market liquidity and trading costs, *Graduate School of Business, Columbia University* .

Keim, Donald B, and Ananth Madhavan, 1998, The cost of institutional equity trades, *Financial Analysts Journal* 50–69.

Ledoit, Olivier, and Michael Wolf, 2004, Honey, i shrunk the sample covariance matrix, *The Journal of Portfolio Management* 30, 110–119.

Ledoit, Olivier, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Ledoit, Olivier, and Michael Wolf, 2011, Robust performances hypothesis testing with the variance, *Wilmott* 55, 86–89.

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

O'Brien, Robert M., 2007, A caution regarding rules of thumb for variance inflation factors, *Quality & Quantity* 41, 673–690.

Patton, Andrew, Dimitris N. Politis, and Halbert White, 2009, Correction to "automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews* 28, 372–375.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Figure 1: Sharpe ratios of the mixed and integrated approach in gray and difference between integrated and mixed approach in green (positive) or red (negative). The analyzed factors are: 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). E.g., the combination of 'value', 'momentum' and 'low volatility' is indicated by VWL. The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). Portfolios are rebalanced monthly from June 1963 to December 2016.

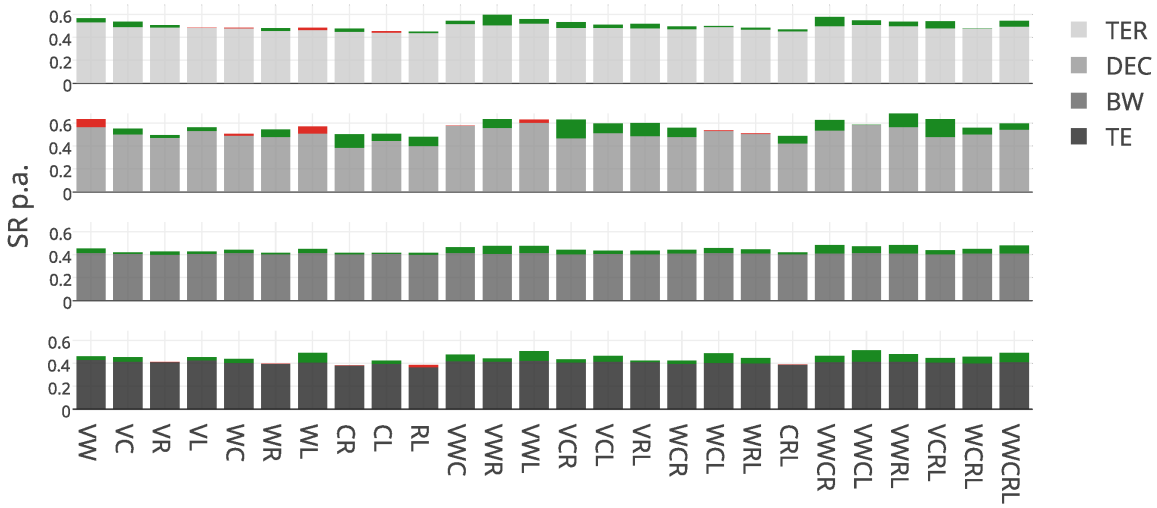

Figure 2: Comparison of the Sharpe ratios of the integrated and mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe ratio (SR diff) in bars and the p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate. The data starts in June 1963 and ends in December 2016.

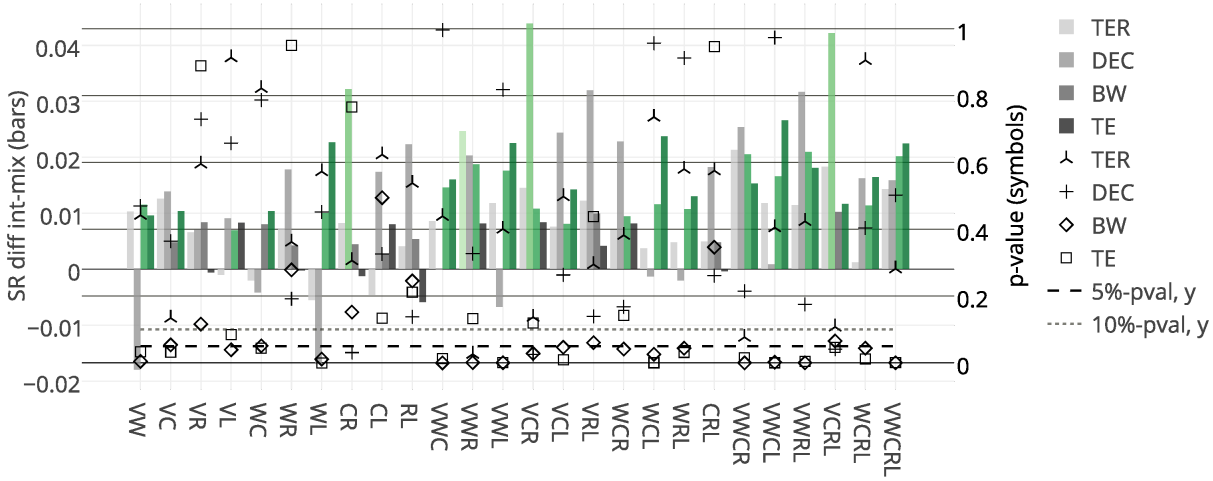

Figure 3: Comparison of the Sharpe ratios of the integrated and mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe ratio (SR diff) in bars and the p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) for a block size of five adjusted by the multiple hypothesis framework of Romano and Wolf (2016) in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate. The data starts in June 1963 and ends in December 2016.

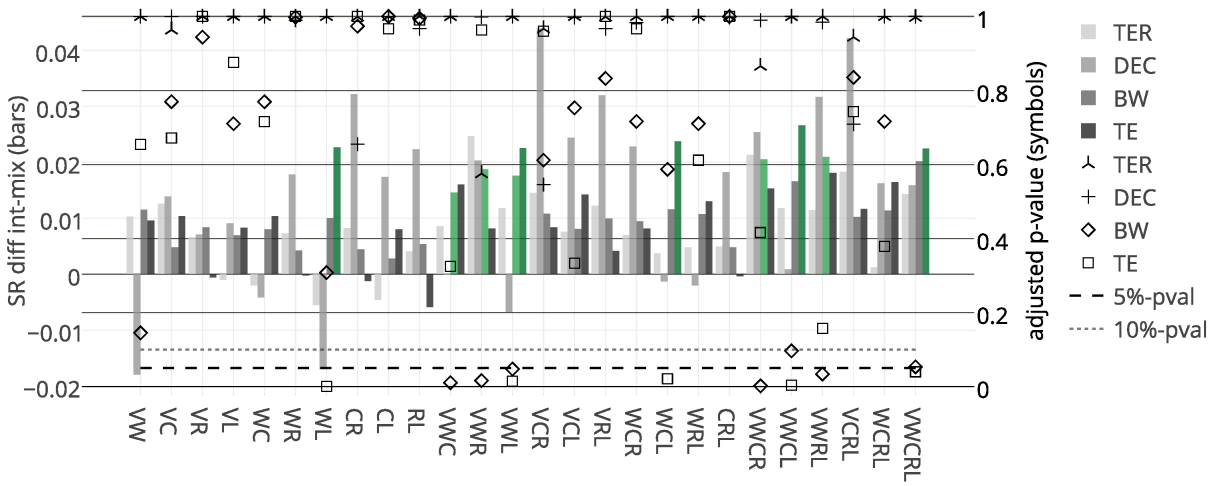

Figure 4: Differences in the logarithmic variance in bars (VR diff) and the adjusted p-values of the robust variance test of Ledoit and Wolf (2011) in symbols in the top chart; differences in the information ratio relative to the market-cap-weighted benchmark (IR diff) in bars and the adjusted p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) in symbols in the middle chart; differences in the logarithmic tracking error in bars (TE diff) and the ajdusted p-values of the robust tracking error test of Ledoit and Wolf (2011) in symbols in the bottom chart. The single hypothesis p-values are adjusted for the number of tries by the multiple hypothesis framework of Romano and Wolf (2016). The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are all four strategies from the set $\mathcal{P}$ in (7). The portfolios are rebalanced monthly from June 1963 to December 2016. The analysis is based on the monthly excess returns above the one-month Treasury bill rate.

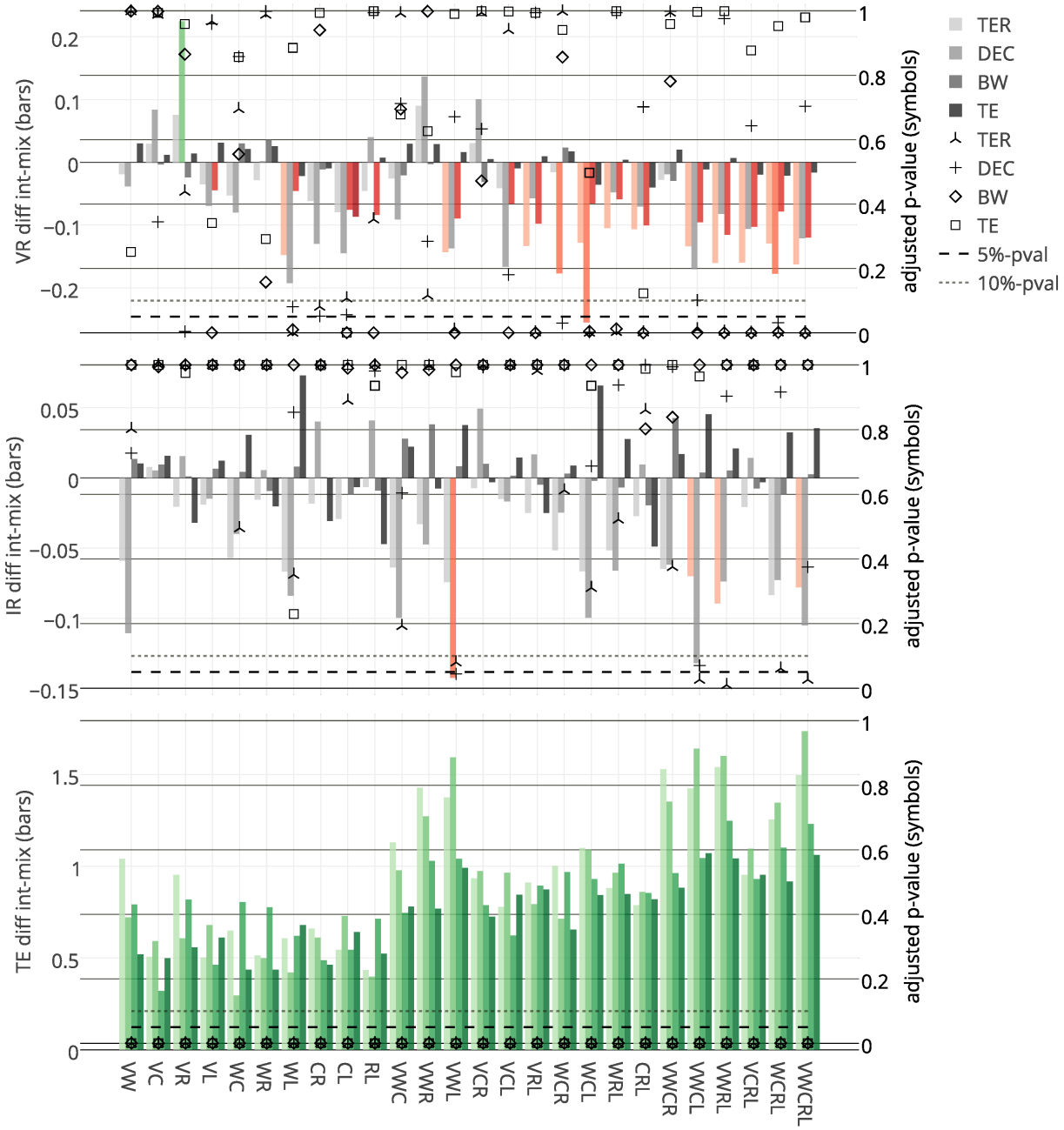

Figure 5: Comparison of the tracking error (TE diff) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The portfolio construction methodologies tested are the TER, BW, and TE. The mixed portfolios is constructed with higher concentrated style portfolios to achieve a similar active risk compared to the integrated approach.The portfolios are rebalanced monthly. We show the difference in the logarithmic tracking error (TE diff) in bars and the p-values of the robust variance test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the market-cap-weighted benchmark. The analysis starts in June 1963 and ends in December 2016.

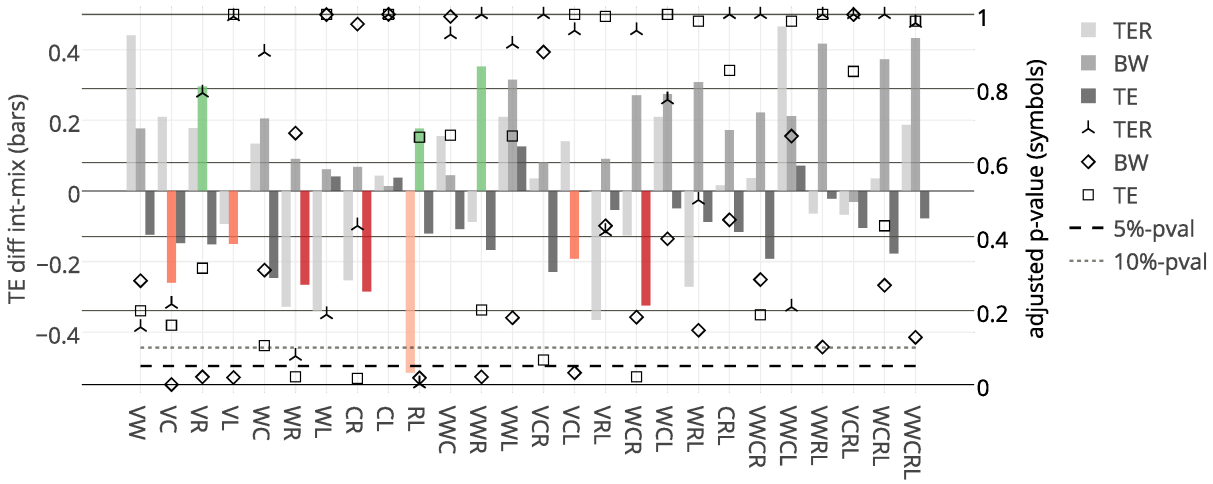

Figure 6: Difference in the Sharpe ratio (SR diff) and the adjusted p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) in the top chart; differences in the logarithmic variance in bars (VR diff) and adjusted p-values of the robust variance test of Ledoit and Wolf (2011) in the middle chart; difference in the information ratio relative to the market-cap-weighted benchmark (IR diff) in bars and the adjusted p-values of the robust Sharpe ratio test in the bottom chart. The single hypothesis p-values are adjusted for the number of tries by the multiple hypothesis framework of Romano and Wolf (2016). The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are TER, BW, and TE. The analysis is based on the monthly excess returns above the market-cap-weighted benchmark. The analysis starts in June 1963 and ends in December 2016.

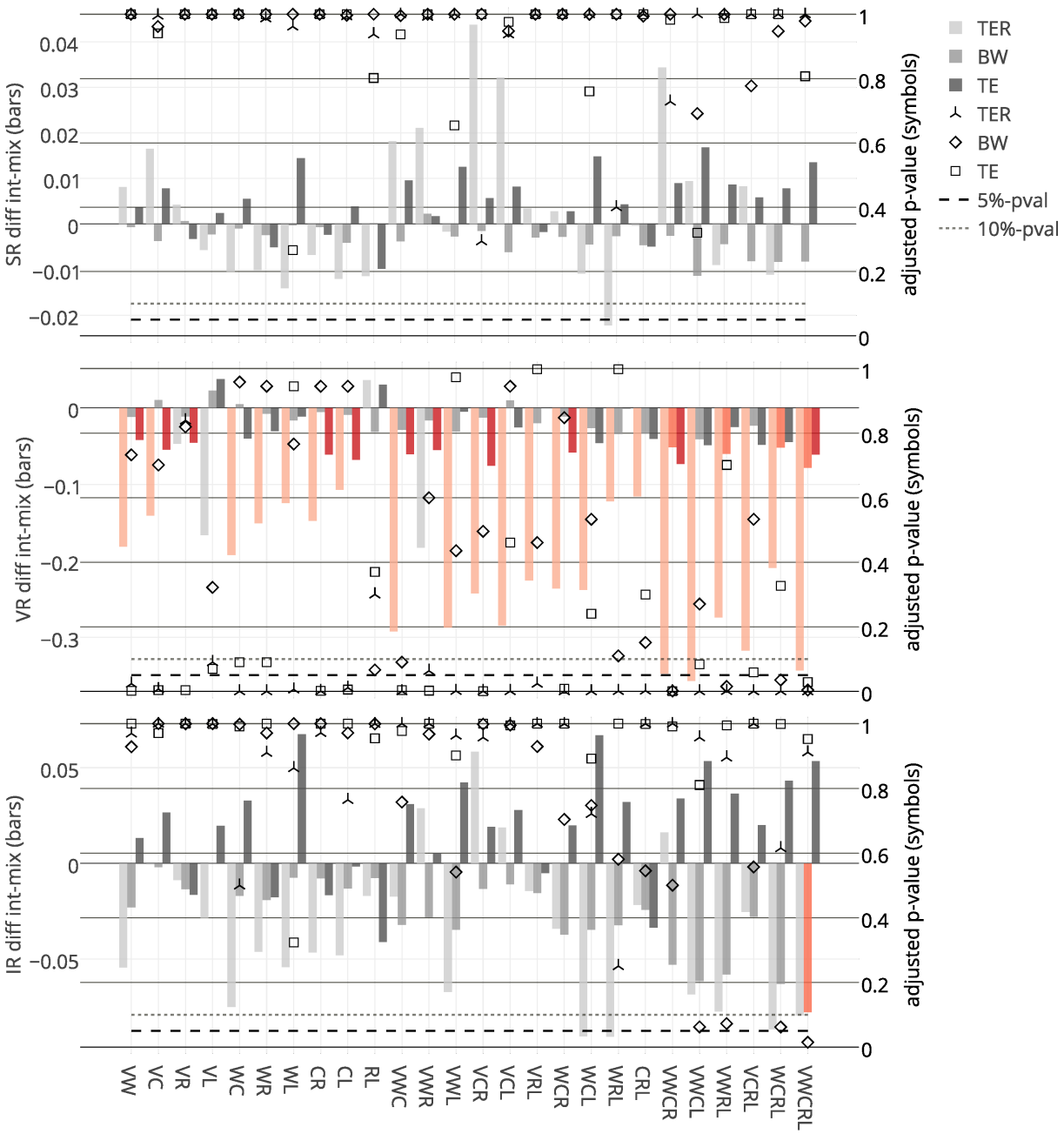

Figure 7: Ordinary least squares regression results with data from June 1963 to December 2016. The dependent variable is the difference in the monthly returns of the integrated approach from those of the mixed approach in long-only style investing. The independent variables are the market portfolio (MKT), the small minus big (SMB), the value (V), the profitability (R), the conservative (C), the momentum (W), and the low volatility (L) factors. Factor returns are calculated by the difference in the performance of the highest to the lowest tercile. Except for the factor small minus big, which is defined as in Fama and French (1992), we only use the big universe to calculate the factor returns. We report the HC3 t-values (t) for the 26 possible factor combinations of V,R,C, W, and L as well as for the tercile (TER), Bender and Wang (2016) (BW), and target tracking error optimization (TE) portfolio construction. HC3 test statistics above (green) and below (red) 1.96 are highlighted and the test statistics are truncated at +/- five for a better overview.

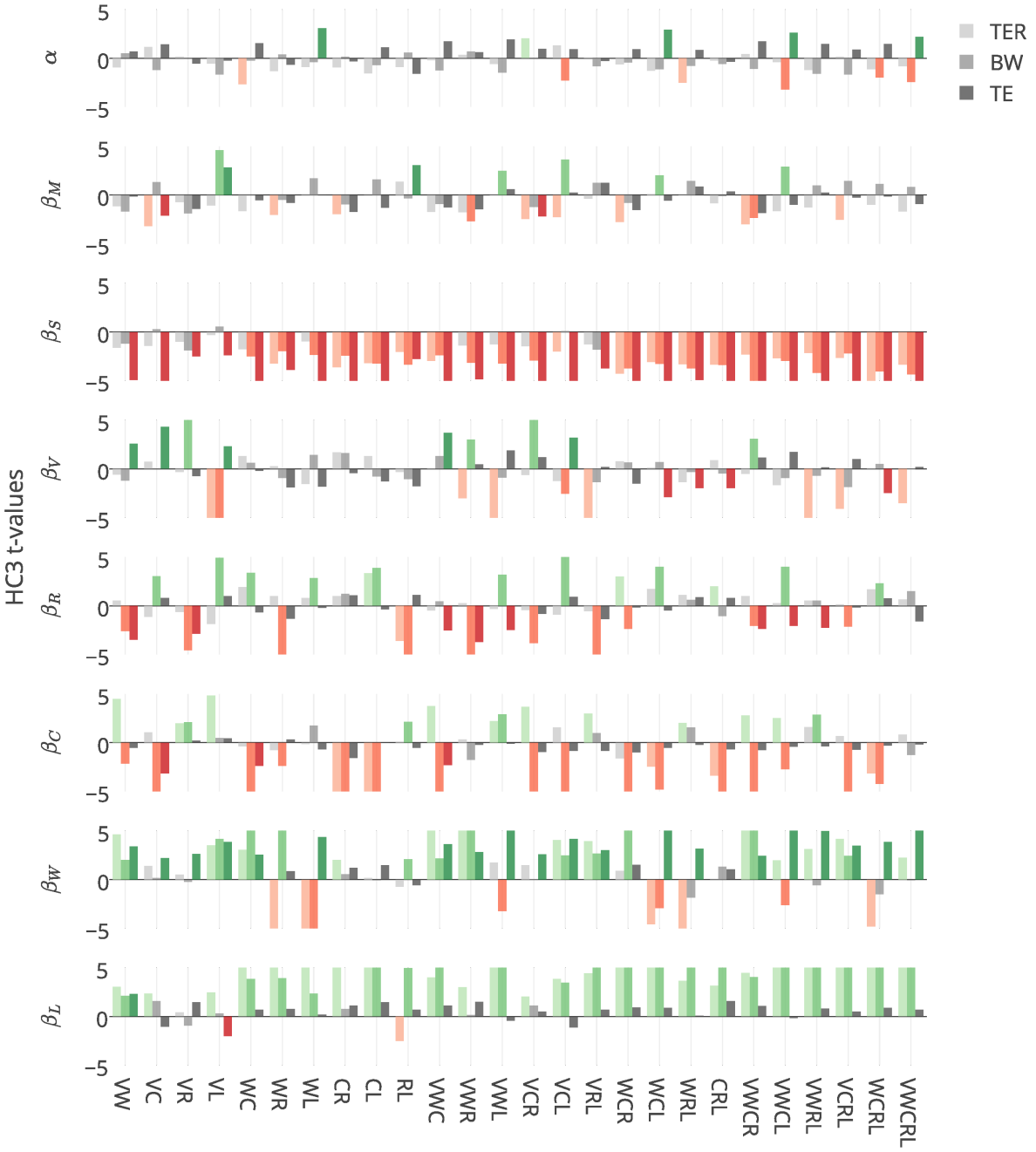

Figure 8: Monthly average turnover of the integrated (integrate), mixed approach with netting (mix - net) and mixed approach without netting (mix - gross) from June 1963 to December 2016. The portfolio construction methodologies tested are the TER, BW, and TE. The portfolios are constructed in a way that the active risk of both, the integrated and mixed approach, are similar over time. A lower (higher) turnover of the integrated approach compared to the mixed approach with netting is highlighted in green (red).

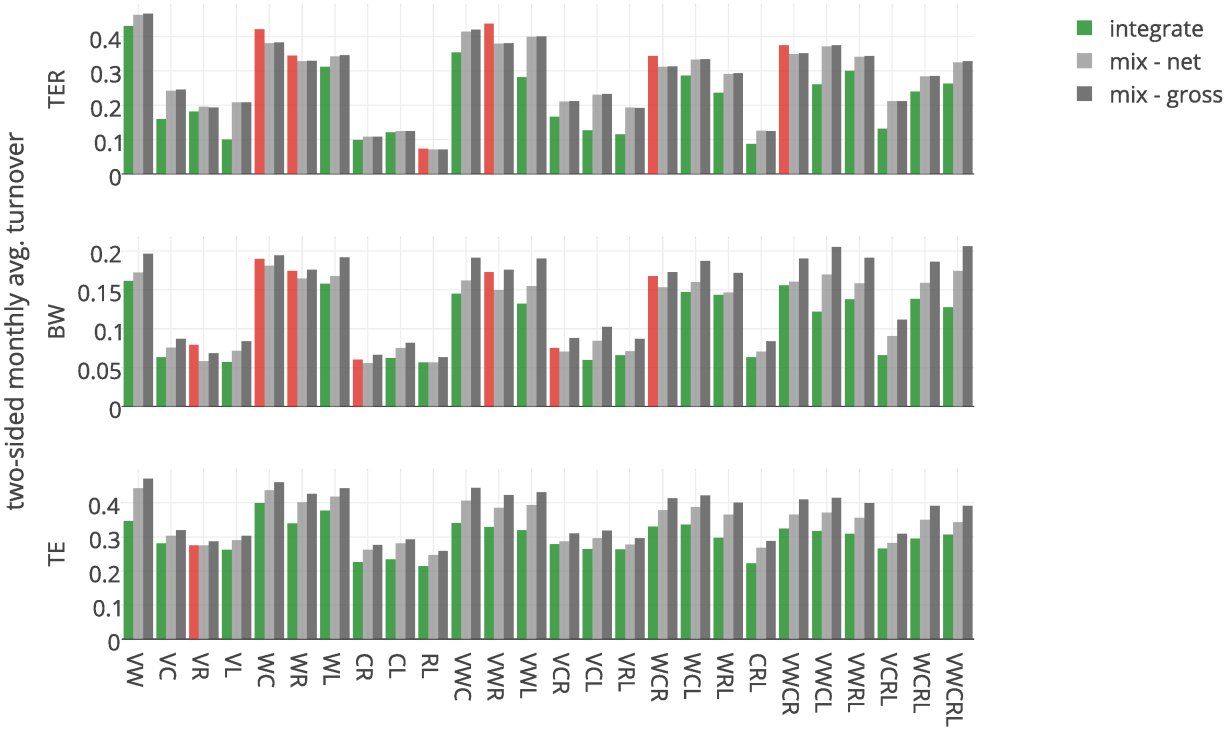

Figure 9: Comparison of the Sharpe (top) and information ratios (bottom) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are TER, BW, and TE. The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe (SR diff) and information ratio (IR diff) in bars and the multiple hypothesis p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate including trading costs. The analysis starts in June 1963 and ends in December 2016.

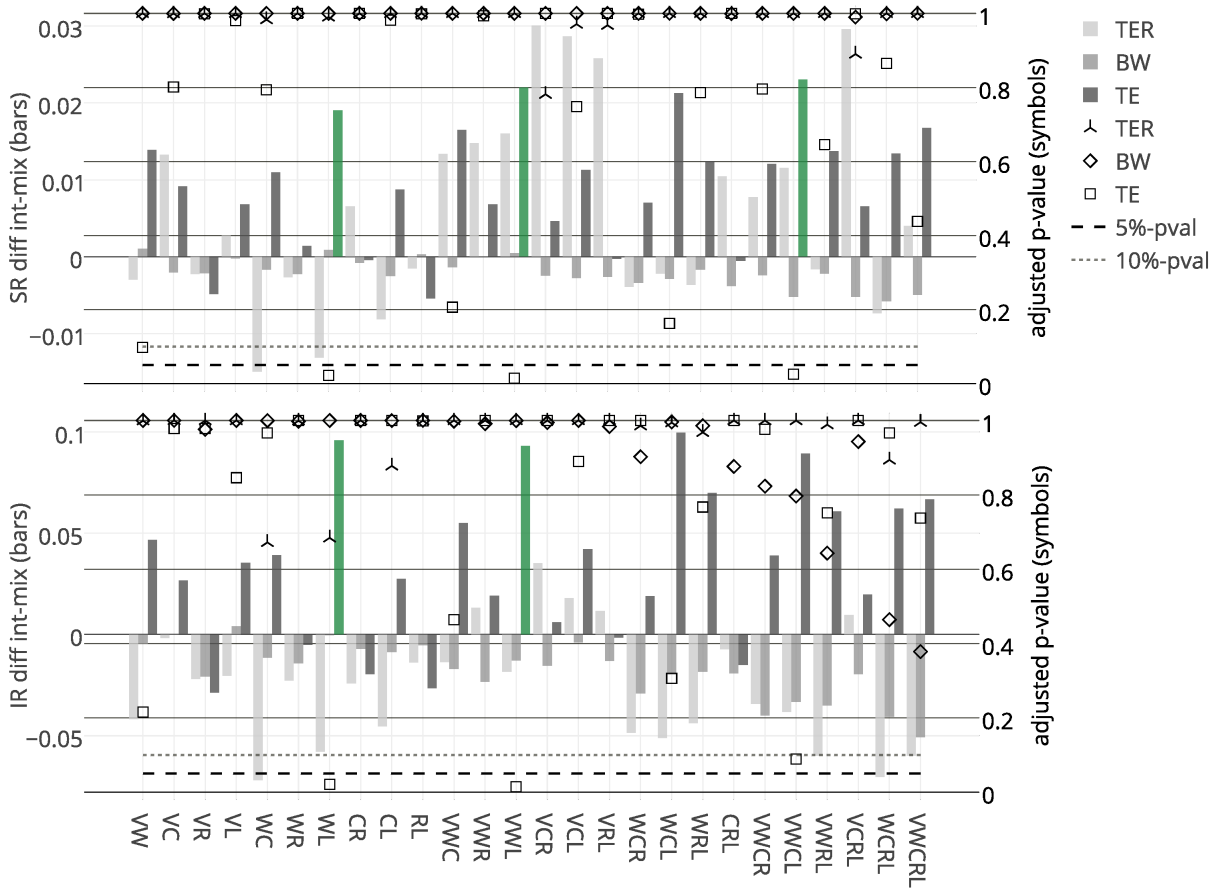

Figure 10: Comparison of the Sharpe (top) and information ratios (bottom) of the integrated approach with those of the mixed approach to long-only style investing. The analyzed factors are 'value' (V), 'momentum' (W), 'investment' (C), 'profitability' (R), and 'low volatility' (L). The portfolio construction methodologies tested are the tercile (TER), Bender and Wang (2016) (BW), and target tracking error optimization (TE). The portfolios are rebalanced monthly. We show the difference in the monthly Sharpe (SR diff) and information ratio (IR diff) in bars and the multiple hypothesis p-values of the robust Sharpe ratio test of Ledoit and Wolf (2008) adjusted by the multiple hypothesis framework of Romano and Wolf (2016) for a block size of five in symbols. The analysis is based on the monthly excess returns above the one-month Treasury bill rate including trading costs. The analysis starts in June 1993 and ends in December 2016.

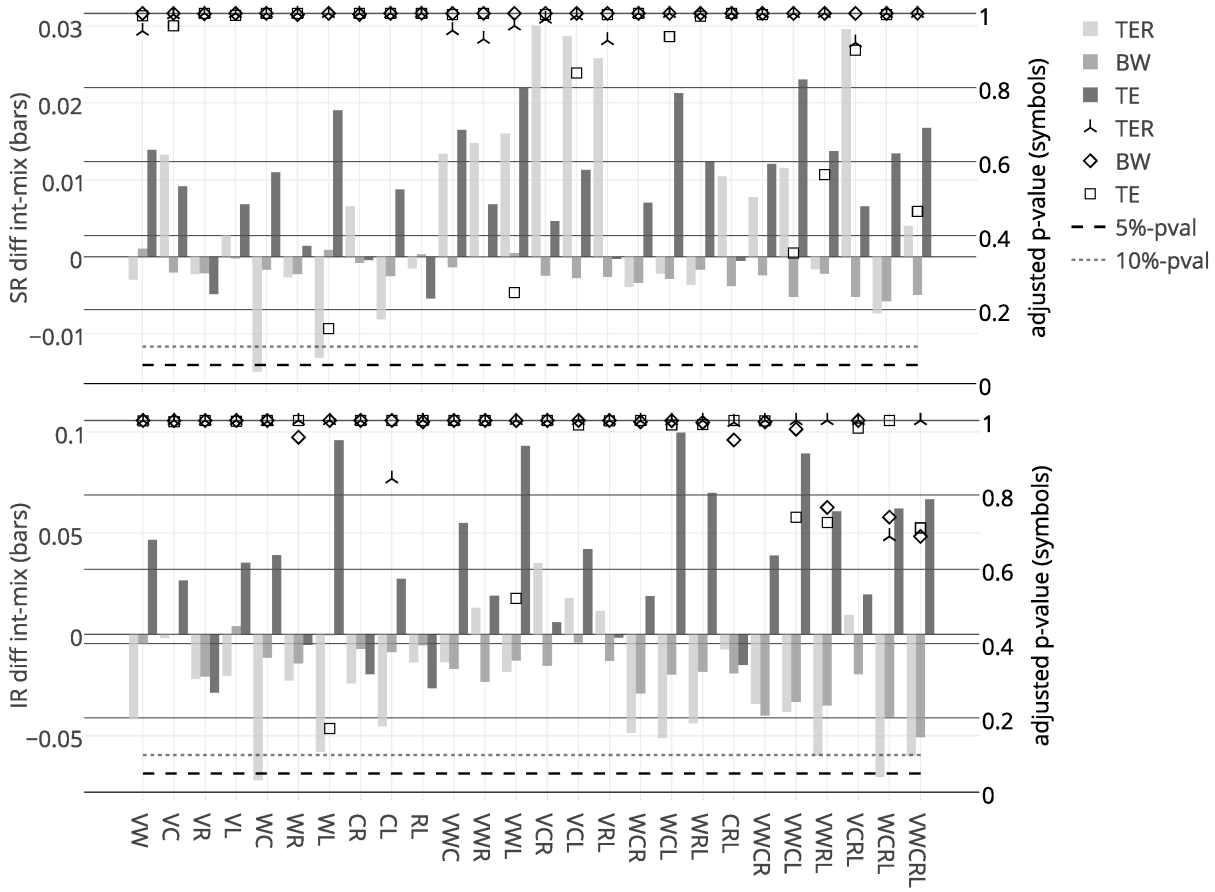

Table 1: Multi-factor ETFs

Most well known multi-factor ETFs as of end of August 16. Data compiled by Bloomberg and ETF.com and ordered by assets under management (AuM) as of end of August 16.

Name	Asset Manager	AuM	Inception	Approach
Goldman Sachs ActiveBeta U.S. Large Cap Equity ETF	Goldman Sachs	$1.15B	01/28/15	mix
FlexShares Morningstar US Market Factor Tilt Index Fund	FlexShares	$842.43M	09/16/11	integrate
John Hancock Multifactor Large Cap ETF	John Hancock	$236.27	09/28/15	integrate
State Street Multi-Factor Global Equity Fund	State Street	$126.39	09/30/14	mix
iShares Edge MSCI Multifactor USA ETF	iShares	$110.03M	04/30/15	integrate
JPMorgan Diversified Return U.S. Equity ETF	JP Morgan	$81.15M	09/29/15	integrate
The Global X Scientific Beta US ETF	Global X	$67.18M	05/12/15	mix
Franklin LibertyQ Global Equity ETF	Franklin	$26.19M	06/01/16	integrate
ETFS Diversified-Factor U.S. Large Cap Index Fund	ETF Securities	$7.82M	01/28/15	mix

Table 2: Stylized example

A stylized example of the calculations of the weights of the mixed and integrated portfolios. We report the market capitalization (mc), the factor values of value (V) and momentum (W), the weights in the factor portfolio of value (w_V) and momentum (w_W), the score for value (s_V), the score for momentum (s_W), the aggregated scores for the combination (s_{VW}), and the resulting weights for the mixed portfolio (w_{mix}) and the integrated portfolio (w_{int}).

	mc	ϕ_V	ϕ_W	s_V^{rank}	s_W^{rank}	$s_{\text{agg}}^{\text{rank}}$	w_V	w_W	w_{mix}	w_{int}
Stock A	10.50	**0.51**	**0.09**	**0.77**	**0.77**	**0.77**	**0.23**	**0.37**	**0.30**	**0.35**
Stock B	5.75	0.22	**0.14**	0.11	**0.88**	0.50	0.00	**0.20**	**0.10**	0.00
Stock C	7.12	0.48	0.02	0.66	0.55	**0.61**	0.00	0.00	0.00	**0.24**
Stock D	22.87	**0.82**	-0.09	**0.88**	0.11	0.50	**0.50**	0.00	**0.25**	0.00
Stock E	7.72	0.32	-0.21	0.44	0.00	0.22	0.00	0.00	0.00	0.00
Stock F	1.15	0.14	0.07	0.00	0.66	0.33	0.00	0.00	0.00	0.00
Stock G	15.72	0.31	0.01	0.33	0.44	0.38	0.00	0.00	0.00	0.00
Stock H	50.91	0.28	-0.07	0.22	0.22	0.22	0.00	0.00	0.00	0.00
Stock I	12.51	**0.97**	**0.22**	**1.00**	**1.00**	**1.00**	**0.27**	**0.43**	**0.35**	**0.42**
Stock J	25.26	0.41	-0.02	0.55	0.33	0.44	0.00	0.00	0.00	0.00

Table 3: Multiple hypothesis testing: Momentum strategy

Comparison of the Sharpe ratios for 20 momentum strategies of IBM compared to the buy and hold strategy. The momentum strategies invest in IBM for the next month if the xth most recent month was positive. Otherwise, we step out of the stock for the next month. We show the strategy for $x = 1, ..., 20$. The out-of-sample backtest starts in June 1963 and ends in December 2014. We show the annualized return (Ret p.a.), the annualized volatility (Vol p.a.), the monthly Sharpe Ratio difference (SR-diff), the bootstrapped p-value of Ledoit and Wolf (2008) with a block size of two, the p-values adjusted by the frameworks of Bonferroni (Bonf), Holm (1979) (Holm), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) (BHY), and Romano and Wolf (2016) (RW). The analysis is based on monthly excess returns above the one-month Treasury bill rate. *** denotes significance at the 0.01 level; ** denotes significance at the 0.05 level; and * denotes significance at the 0.1 level.

x =	1	2	3	4	5	6	7	8	9	10
Ret p.a.	0.051	0.023	0.021	0.008	0.010	-0.008	0.023	0.012	0.041	0.029
Vol p.a.	0.163	0.164	0.158	0.159	0.160	0.176	0.173	0.160	0.166	0.175
SR-diff	0.030	-0.018	-0.021	-0.044	-0.041	-0.069	-0.019	-0.037	0.012	-0.010
pval	0.347	0.576	0.507	0.170	0.203	0.015**	0.511	0.242	0.708	0.742
Bonf	1.000	1.000	1.000	1.000	1.000	0.292	1.000	1.000	1.000	1.000
Holm	1.000	1.000	1.000	1.000	1.000	0.277	1.000	1.000	1.000	1.000
BHY	0.645	0.752	0.731	0.485	0.506	0.146	0.731	0.537	0.752	0.752
RW	0.985	0.994	0.994	0.904	0.933	0.230	0.994	0.949	0.996	0.996

x =	11	12	13	14	15	16	17	18	19	20
Ret p.a.	0.006	0.042	0.004	0.019	-0.013	0.071	0.040	0.026	0.017	0.005
Vol p.a.	0.171	0.171	0.163	0.159	0.175	0.165	0.176	0.163	0.168	0.164
SR-diff	-0.047	0.012	-0.051	-0.026	-0.077	0.062	0.009	-0.013	-0.028	-0.049
pval	0.127	0.681	0.100	0.449	0.008***	0.053*	0.752	0.691	0.355	0.138
Bonf	1.000	1.000	1.000	1.000	0.168	1.000	1.000	1.000	1.000	1.000
Holm	1.000	1.000	1.000	1.000	0.168	0.958	1.000	1.000	1.000	1.000
BHY	0.461	0.752	0.461	0.731	0.146	0.355	0.752	0.752	0.645	0.461
RW	0.848	0.996	0.818	0.993	0.141	0.580	0.996	0.996	0.985	0.857

Table 4: Factors' summary statistics

Annualized return (Ret p.a.), annualized volatility (Vol p.a.), Sharpe ratio (SR p.a.), and maximum drawdown (Max. Draw.) for the value-weighted monthly excess returns above the one-month Treasury bill rate. The period starts in June 1963 and ends in December 2016. The market includes all securities. Small: the securities below the NYSE market capitalization median; Big: the securities above the NYSE market capitalization. The other factors are the top (first-listed), respectively, bottom (second-listed), terciles in the big universe of securities of the following factors: 'value' (Value - Growth), 'robustness' (Robust - Weak), 'investment' (Conser. - Aggr), 'momentum' (Winner - Loser), and 'low volatility' (LowVol - HighVol).

1963 - 2016	Market	Small	Big	Value	Growth	Robust	Weak	Conser.	Aggr.	Winner	Loser	LowVol	HighVol
Ret p.a.	5.14	6.48	5.07	8.67	4.71	6.39	2.76	7.11	4.90	8.10	2.39	5.50	4.49
Vol p.a.	15.53	21.55	15.13	16.38	16.19	15.22	18.11	15.06	17.90	16.91	19.81	12.70	25.67
SR p.a.	0.33	0.30	0.34	0.53	0.29	0.42	0.15	0.47	0.27	0.48	0.12	0.43	0.17
Max. Draw.	55.23	73.83	55.44	54.11	58.60	53.04	78.59	47.54	63.72	50.39	72.71	48.90	78.84

Table 5: Variance inflation factors

Variance inflation factors of the following factors: value (V), momentum (W), investment (C), profitability (R), and low volatility (L). The return series of each factor are computed as the value-weighted return of the upper tercile less the value-weighted return of the lower tercile of each factor for the big universe. We show the variance inflation factors for each of the 5 factors as independent variables. The dependent variable of the model is shown on the horizontal axis, while the independent variables are shown on the vertical axis.

	V	C	W	R	L
V	-	1.49	1.03	1.36	1.73
C	1.46	-	1.16	1.38	1.48
W	1.74	1.99	-	1.42	1.76
R	1.91	1.97	1.18	-	1.32
L	1.95	1.70	1.18	1.06	-