Imperial College London

PORTFOLIO MANAGEMENT RESEARCH PAPER

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

The Mixed vs Integrated Approach to Style Investing

Author:
Adan Fhima (CID 02226023)
Chedi Mnif (CID 06064588)
Christopher Dybdahl (CID 06060098)

Date: December 18, 2025

Abstract

This paper revisits and extends the analysis of Leippold et al. [2018] on the comparison between mixed and integrated approaches to multi-factor investing. Using a comprehensive dataset of U.S. equities from 1963 to 2024, we construct 104 portfolios for both the integrated and mixed approaches by applying four portfolio formation rules to 26 combinations of five factors, and evaluate their performance using robust hypothesis testing. We find that the integrated approach exhibits statistically significant higher Sharpe ratios under the bootstrap tests of Ledoit and Wolf [2008, 2011] for many portfolio compositions. However, these differences are no longer statistically significant once multiple hypotheses are adjusted for using the stepdown procedure of Romano et al. [2016]. Leippold et al. [2018] argues that the integrated and mixed approaches generate portfolios with significantly different active risk even when portfolio formation rules are identical. Motivated by this, we repeat the robust tests after modifying the portfolio formation rules to promote equal active risk, and find that the integrated approach yields some, but now fewer, statistically significant improvements in Sharpe ratios. As before, these differences are not statistically significant after applying the stepdown procedure. Overall, our results confirm that the integrated approach does not reliably outperform once multiple hypothesis testing is properly accounted for.

Contents

1 Introduction

Practitioners and academics broadly agree that factors such as value, momentum, and low volatility command risk premiums, yet surprisingly little consensus exists on how these signals should be combined within a single portfolio. With global assets in smart beta exchange-traded funds surpassing $1.56 trillion as of early 2024 ETFGI [2024], the question of how to optimally harvest factor premiums carries substantial practical consequences. Among the most debated implementation choices is the method by which multiple factors should be combined within a single portfolio. Two fundamentally different philosophies have emerged: the *mixed approach*, which constructs separate portfolios for each factor and subsequently combines them, and the *integrated approach*, which first aggregates factor signals at the security level before forming a single portfolio. Despite the considerable capital allocated to strategies employing each method, surprisingly little consensus exists regarding their relative merits.

This study revisits and extends the analysis of Leippold et al. [2018], who provide a rigorous comparison of these competing approaches. Our investigation centers on three key findings that challenge conventional wisdom in factor investing.

First, we examine what appears to be a violation of the fundamental risk-return tradeoff. Several influential studies, including Fitzgibbons et al. [2016] and Clarke et al. [2016], suggest that the integrated approach delivers superior risk-adjusted returns compared to the mixed approach. If true, this would constitute a rare "free lunch" in financial markets: higher expected returns accompanied by lower volatility. We investigate whether this apparent anomaly reflects a genuine inefficiency or merely a statistical artifact arising from methodological limitations in prior research.

Second, we highlight the critical importance of robust statistical inference when evaluating investment strategies. The literature on factor investing suffers from a well-documented multiple testing problem [Harvey et al., 2016]. When researchers examine numerous strategy variations, some will inevitably appear statistically significant by chance alone. Bailey et al. [2014] demonstrate that without proper adjustment for multiple comparisons, backtested strategies frequently fail to deliver their promised performance out-of-sample. By applying the stepdown procedure of Romano and Wolf [2005], which controls the family-wise error rate[1] while accounting for dependence among test statistics, we reassess whether the integrated approach's apparent outperformance survives rigorous testing.

Third, we uncover a mechanism that explains the observed performance differences between the two approaches. The integrated method, by construction, tends to exclude securities with extreme factor exposures. A stock ranking poorly on one factor but excellently on another will receive a moderate aggregated score, potentially falling outside the portfolio's selection threshold. This implicit screening creates an unintended tilt toward lower-volatility securities, as extreme stocks whether extremely cheap, extremely expensive, or exhibiting extreme momentum, tend to

[1] The family-wise error rate (FWER) is the probability of making at least one Type I error (a false positive) when performing multiple statistical tests.

be more volatile. Consequently, the integrated approach embeds exposure to the low-volatility anomaly documented by Blitz and van Vliet [2007], which explains its lower risk profile but does not constitute genuine alpha generation.

The practical relevance of this comparison extends beyond academic interest. As Leippold et al. [2018] document, the asset management industry remains divided on implementation methodology. Major providers such as Goldman Sachs and J.P. Morgan favor the mixed approach in their multi-factor offerings, while competitors including Research Affiliates and AQR advocate for integrated strategies. Given the magnitude of assets following these competing philosophies, even small differences in expected returns translate to economically significant implications for investor wealth.

From a methodological perspective, our analysis contributes to the growing literature emphasizing the importance of proper statistical inference in empirical finance. The techniques we employ, particularly the Ledoit and Wolf [2008] test for Sharpe ratio comparisons and the Romano et al. [2016] procedure for multiple hypothesis adjustment, represent current best practices for strategy evaluation and preventing false discoveries.

Implementing this comparison presents several challenges that merit acknowledgment. The analysis requires constructing portfolios across five distinct factors: value, momentum, low volatility, profitability, and investment - yielding 26 possible factor combinations when considering all subsets of two or more factors. Each combination must be evaluated under multiple portfolio construction methodologies, and the resulting performance differences subjected to statistical tests that account for both time-series dependence and multiple comparisons. The computational burden is non-trivial, and the data requirements span decades of both return and balance sheet information. Consequently, in our baseline analysis we employ three portfolio formation rules. In a subsequent analysis which aims to control for same active risk, we additionally consider a fourth portfolio formation rule based on tracking error.

Our contribution is threefold. We replicate the original study of Leippold et al. [2018] using an extended sample period through 2024, providing out-of-sample evidence on the stability of their conclusions. We offer an accessible yet rigorous treatment of the robust testing framework, making these valuable techniques more approachable for practitioners and researchers. Finally, we document any deviations from the original findings that emerge in the more recent data, contributing to our understanding of how factor investing dynamics may have evolved.

The remainder of this paper proceeds as follows. Section 2 develops the theoretical framework, presenting both the portfolio construction methodologies and the statistical testing procedures employed. Section 3 describes our data sources, factor definitions, and implementation details. Section 4 presents the empirical findings, including performance comparisons, statistical significance tests, and analysis of factor exposures. Section 4.3 concludes with a discussion of practical implications and directions for future research.

2 The Robust Testing Framework

A meaningful comparison between the integrated and mixed approaches requires statistical procedures that remain valid under heteroskedasticity, serial correlation, and extensive multiple testing. Standard Gaussian asymptotics are inappropriate in this setting because portfolio returns display persistent autocorrelation and because evaluating dozens of factor combinations induces severe data-snooping risk. We therefore rely on the robust Sharpe-ratio methodology of Ledoit and Wolf [2008, 2011], complemented by the resampling-based multiple-testing adjustment of Romano et al. [2016].

2.1 Sharpe Ratio Difference Testing

Let r_t^{int} and r_t^{mix} denote the excess returns of the integrated and mixed portfolios, and define

$$d_t = r_t^{\text{int}} - r_t^{\text{mix}}.$$

The parameter of interest is the directional Sharpe-ratio difference

$$\Delta = SR_{\text{int}} - SR_{\text{mix}},$$

under the hypotheses

$$H_0 : \Delta = 0, \qquad H_1 : \Delta > 0.$$

Following Ledoit and Wolf [2008], the Sharpe ratio is a smooth transformation of the first two unconditional moments. Define

$$\nu = (\mu_i, \mu_n, \gamma_i, \gamma_n)'$$

with means $\mu_i = \mathbb{E}(r_t^{\text{int}})$, $\mu_n = \mathbb{E}(r_t^{\text{mix}})$ and second moments $\gamma_i = \mathbb{E}(r_t^{\text{int2}})$, $\gamma_n = \mathbb{E}(r_t^{\text{mix2}})$. Let $\hat{\nu}$ denote sample estimates. Then

$$\Delta = f(\nu), \qquad \hat{\Delta} = f(\hat{\nu}),$$

with

$$f(a, b, c, d) = \frac{a}{\sqrt{c - a^2}} - \frac{b}{\sqrt{d - b^2}}. \tag{1}$$

Under weak dependence and finite fourth moments,

$$\sqrt{T}\,(\hat{\nu} - \nu) \xrightarrow{d} N(0, \Psi), \tag{2}$$

and the delta method yields

$$\sqrt{T}(\hat{\Delta} - \Delta) \xrightarrow{d} N(0,\ \nabla f(\nu)'\Psi \nabla f(\nu)). \tag{3}$$

The gradient of f is

$$\nabla f(a,b,c,d) = \begin{pmatrix} \frac{c}{(c-a^2)^{3/2}} \\ -\frac{d}{(d-b^2)^{3/2}} \\ \frac{a}{2(c-a^2)^{3/2}} \\ \frac{b}{2(d-b^2)^{3/2}} \end{pmatrix}. \tag{4}$$

The long-run covariance matrix Ψ is heteroscedasticity and autocorrelation robust (HAC). Define

$$y_t = \left(r_t^{\text{int}} - \mu_i,\ r_t^{\text{mix}} - \mu_n,\ r_t^{\text{int2}} - \gamma_i,\ r_t^{\text{mix2}} - \gamma_n\right)'.$$

Then

$$\Psi = \lim_{T\to\infty} \sum_{j=-(T-1)}^{T-1} \Gamma_T(j), \tag{5}$$

where $\Gamma_T(j)$ is the empirical autocovariance at lag j. We estimate Ψ using the prewhitened QS kernel of Andrews and Monahan [1992], as recommended by Ledoit and Wolf [2008].

Because asymptotic approximations perform poorly in finite samples, Ledoit and Wolf [2008] propose a studentised circular block bootstrap. For each bootstrap replication we compute

$$\hat{v}^*, \qquad \hat{\Psi}^* = \frac{1}{\ell}\sum_{j=1}^{\ell} \zeta_j^* \zeta_j^{*\prime}, \qquad \hat{\Delta}^* = f(\hat{v}^*), \tag{6}$$

where ζ_j^* are the block-aggregated moment observations

$$\zeta_j = \frac{1}{\sqrt{b}} \sum_{t=1}^{b} y^*_{(j-1)b+t}, \quad t = 1,\ldots,l.$$

Here, the parameter b denotes the block size in the bootstrap and is calibrated from the data. Following Ledoit and Wolf [2011], we use the value $b = 5$ resulting from their calibration procedure.

The bootstrap standard error is

$$s(\hat{\Delta}) = \sqrt{\frac{\nabla f(\hat{v})' \hat{\Psi} \nabla f(\hat{v})}{T}}. \tag{7}$$

The original test statistic is the two-sided measure

$$d = \frac{|\hat{\Delta}|}{s(\hat{\Delta})}, \tag{8}$$

with bootstrap analogue

$$d^{*,m} = \frac{|\hat{\Delta}^{*,m} - \hat{\Delta}|}{s(\hat{\Delta}^{*,m})}, \tag{9}$$

and p-value

$$p = \frac{\#\{d^{*,m} \geq d\} + 1}{M + 1}. \tag{10}$$

Because our economic question is directional, taking absolute values is inappropriate: large negative $\hat{\Delta}$ would falsely appear significant. We therefore use the signed statistic

$$T = \frac{\hat{\Delta}}{s(\hat{\Delta})}, \qquad T^{*,m} = \frac{\hat{\Delta}^{*,m} - \hat{\Delta}}{s(\hat{\Delta}^{*,m})}, \tag{11}$$

and compute the one-sided p-value

$$p = \frac{\#\{T^{*,m} \geq T\} + 1}{M + 1}. \tag{12}$$

2.2 Information Ratio, Variance, and Tracking-Error

For completeness, we also report differences in information ratios, variances, and tracking errors. However, to keep the focus of the paper concise, we do not conduct hypothesis tests for information ratio and variance differences, as in Leippold et al. [2018], and instead present these measures descriptively.

2.3 Multiple-Hypothesis Adjustment

Assessing twenty-six factor combinations per portfolio construction rule introduces a heightened risk of spurious significance arising from data snooping and multiple testing. Unadjusted p-values would therefore yield inflated false discoveries. We employ the Romano–Wolf stepdown procedure of Romano et al. [2016], which delivers dependence-robust adjusted p-values.

Let $T_{(s)}$ denote the ordered Sharpe-ratio statistics (from most to least significant). The initial bootstrap p-value is

$$\hat{p}^{\text{init}}_{(s)} = \frac{\#\{\max_{j \geq s} T^{*,m}_{(j)} \geq T_{(s)}\} + 1}{M + 1}. \tag{13}$$

To enforce monotonicity (Remark 4.1 in Romano et al. [2016]),

$$\hat{p}^{\text{adj}}_{(1)} = \hat{p}^{\text{init}}_{(1)}, \qquad \hat{p}^{\text{adj}}_{(s)} = \max\left(\hat{p}^{\text{adj}}_{(s-1)},\ \hat{p}^{\text{init}}_{(s)}\right). \tag{14}$$

A Sharpe-ratio improvement is deemed statistically meaningful only if it survives both the robust time-series test and the Romano–Wolf multiplicity correction. This dual requirement follows the best-practice recommendations of Bailey et al. [2014] and Harvey et al. [2016].

3 Data and Methodology

3.1 Data Sources

We construct our dataset by combining information from three databases accessed via the Wharton Research Data Services (WRDS) platform. Compustat provides annual income statement and balance sheet data, which are used to construct the fundamental factor characteristics. The Center for Research in Security Prices (CRSP) supplies monthly stock returns, which enter both the construction of the momentum factor and the portfolio backtests. Finally, monthly risk-free rates are obtained from the CRSP Treasury file and converted as follows

$$r_f = (1 + y/100)^{1/12} - 1,$$

where y denotes the annualised 30-day T-bill yield. All series span the period 1962 to 2024, as Compustat data prior to this period are subject to serious selection bias toward historically large and successful firms [Fama and French, 1992]. To ensure that accounting information is available at the time of portfolio construction, we lag the annual Compustat data by six months.

We apply the filters of Leippold et al. [2018]. The sample includes only common shares (SHRCD 10 and 11) listed on NYSE, AMEX, or NASDAQ, and excludes finance, insurance, and real estate firms (SIC 6000–6799). These firms are omitted because high leverage is a structural feature of their business models, whereas similar leverage levels would typically signal financial distress for nonfinancial firms [Fama and French, 1992].

In historical portfolio analyses, excluding firms with missing observations or discarding periods with incomplete data can induce selection and lookahead biases, as such filters implicitly condition on future performance and data availability, thereby favoring stronger firms or more favorable periods. To mitigate these biases, we define the investable universe using a pre-specified rule based exclusively on information available at the time. Importantly, once specified, this rule is held fixed, as adjusting it ex post can itself introduce lookahead bias. Specifically, we compute the NYSE median market capitalisation each December and retain firms above this threshold.

After filtering, we compute the five factor characteristics. Some characteristics, such as investment, require lagged balance sheet information and therefore generate missing values. The corresponding firm–year observations are excluded, as this missingness is purely mechanical and does not introduce bias. We then remove firms with incomplete factor information, following the methodology of Leippold et al. [2018]. This step results in only 84 firm deletions over the entire sample period within the filtered universe, which we assess as inducing negligible bias. Updating the universe filter ex post is avoided to prevent the introduction of lookahead bias.

The final dataset contains 3,670 firms, 41,933 firm-year observations, and approximately 980,000 firm-month returns covering 756 months from 1963 to 2024. The cross-section averages roughly 700 stocks per year, ranging from 378 in 1964 to 1,045 at the height of the dot-com boom.

3.2 Factor Definitions

We construct five characteristics following Leippold et al. [2018] and drawing on the definitions of Fama and French [2015]. Each factor is measured annually at calendar year-end, and all are oriented so that higher values indicate more desirable exposure.

Value (V). Defined as the inverse of the price-to-book ratio. Book equity is computed as common equity plus deferred taxes minus preferred stock; when unavailable, we use total assets minus total liabilities. Firms with zero or negative book equity are excluded. Lower price-to-book implies stronger value exposure.

Momentum (W). Measured as the geometric mean of monthly returns over the calendar year:

$$W_i = \left(\prod_{m=1}^{12} (1 + r_{i,m}) \right)^{1/12} - 1.$$

Low Volatility (L). Computed as the 36-month rolling standard deviation of monthly returns, requiring at least 12 observations. Because lower volatility is desirable, we negate the measure before scoring so that stable firms receive higher scores.

Profitability (R). Defined as operating income relative to book equity:

$$R_i = \frac{\text{OIBDP}_i - \text{DP}_i}{\text{Book Equity}_i},$$

capturing the efficiency with which firms generate earnings from their balance sheet.

Investment (C). Given by

$$C_i = \frac{\text{Total Assets}_{t-1}}{\text{Total Assets}_t},$$

so that firms expanding aggressively have $C < 1$, while conservative firms have $C \geq 1$. This definition removes the first year-end observation for each firm due to the lag requirement.

All characteristics are winsorised at the 0.5th and 99.5th percentiles cross-sectionally each year. Because portfolio formation is primarily rank-based, winsorisation affects only extreme outliers and does not materially influence scores.

3.3 Implementation Details

The full comparison spans 208 portfolios: each of the 26 factor combinations is implemented under both the mixed and integrated approaches using four portfolio formation rules.

The first two rules are selection-based. The tercile method (TER) ranks stocks by their factor score and selects the top third; the decile method (DEC) selects the top tenth. Within each group, stocks are weighted by market capitalisation. These constructions follow the style of Fama and French [1992].

The remaining two rules are benchmark-oriented. The Bender–Wang method (BW), introduced by Bender and Wang [2016], starts from market-capitalisation weights and applies multiplicative adjustments based on factor scores: stocks with high scores receive larger weights, stocks with low scores receive smaller weights, but all stocks remain in the portfolio. The tracking-error method (TE), following Fitzgibbons et al. [2016], solves an optimisation that maximises factor exposure subject to a bound on deviation from the market benchmark. We estimate the covariance matrix via Ledoit–Wolf shrinkage on 36-month returns, re-estimated each December.

Portfolios are formed annually. Factor scores are computed at year-end, weights take effect the following January, and during the year they drift with returns and are renormalised monthly to maintain full investment, consistent with Leippold et al. [2018].

Inference follows the bootstrap procedure of Section 2.1. We resample returns in five-month blocks, draw 1,000 replications per test, and apply Romano–Wolf adjustments to control false discoveries across the 26 hypotheses per construction rule.

Table 1 reports variance inflation factors (VIFs) for using the TER portfolio returns for each of the five factors. Panel A presents standard VIFs obtained by regressing each factor on the remaining four factors. Panel B reports pairwise VIFs, included for comparability with the original study. Across both panels, VIF values remain below the threshold of 5 used in Leippold et al. [2018], indicating limited multicollinearity.

Panel A: VIF all factors jointly

Factor	VIF
V	3.94
W	2.32
C	3.47
R	2.07
L	2.77

Panel B: Pairwise VIF diagnostics

Factor pair	VIF
V-W	1.12
V-C	1.89
V-R	1.24
V-L	1.13
W-C	1.00
W-R	1.02
W-L	1.00
C-R	1.03
C-L	1.71
R-L	1.14

Table 1: Variance inflation factor diagnostics. Panel A reports standard variance inflation factors computed by regressing each factor on the remaining four factors. Panel B reports VIFs from two-regressor auxiliary regressions for each factor pair.

4 Empirical Results

This section presents our empirical findings. We begin with an overview of portfolio performance across the 26 factor combinations, then examine whether observed differences between the mixed and integrated approaches are statistically significant. We report results both with and without adjustment for multiple hypothesis testing.

4.1 Equal portfolio formation rules

Figure 1 reports annualised Sharpe ratios for all 26 factor combinations under the tercile, decile, and Bender–Wang methods. The grey bar marks the lower Sharpe ratio between integrated and mixed; the coloured tip shows the gap, green when integrated wins, red when mixed wins.

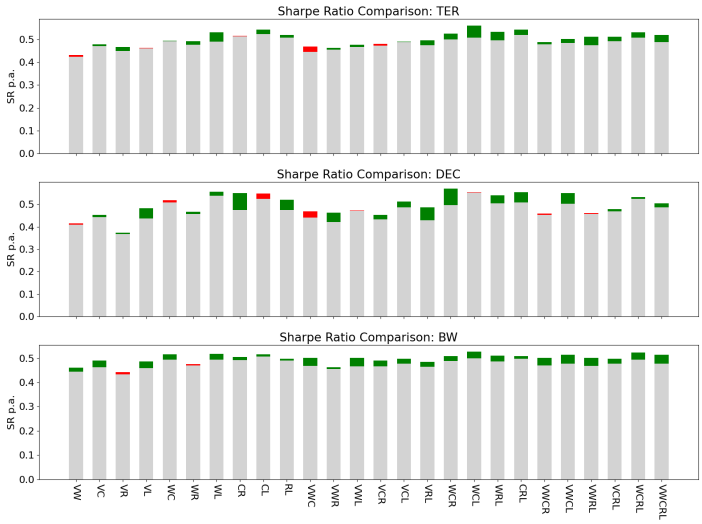


Figure 1: Sharpe ratios for all factor combinations. Grey bars show the minimum of integrated and mixed; green (red) tips indicate integrated (mixed) outperformance.

Sharpe ratios range from 0.4 to 0.55 across specifications. The integrated approach delivers higher risk-adjusted returns in most cases, especially when the low-volatility factor is included. This is not surprising as under the integrated method, a stock must score well on average across all factors to enter the portfolio. Stocks with a strong tilt toward a single factor but poor exposure elsewhere are filtered out. What remains are stocks with balanced characteristics, and these tend to be less volatile. The Bender-Wang method shows smaller differences between the two approaches because it assigns weights to all stocks rather than selecting a subset, so both approaches produce similar portfolios.

4.1.1 Sharpe Ratio

The Sharpe ratio gaps documented above could be genuine or could be noise. We apply the bootstrap test of Ledoit and Wolf [2008] to each of the 26 factor combi-

nations. Figure 2 plots the Sharpe ratio difference (integrated minus mixed) as bars and the corresponding p-value as symbols. Positive bars mean integrated outperforms; negative bars favour mixed. The dashed and dotted lines mark the 5% and 10% significance levels.

Most differences are positive but small, below 0.05 in annual Sharpe ratio units. Only a handful reach conventional significance when tested one at a time. Tercile and decile methods produce the largest gaps; Bender-Wang differences hover near zero.

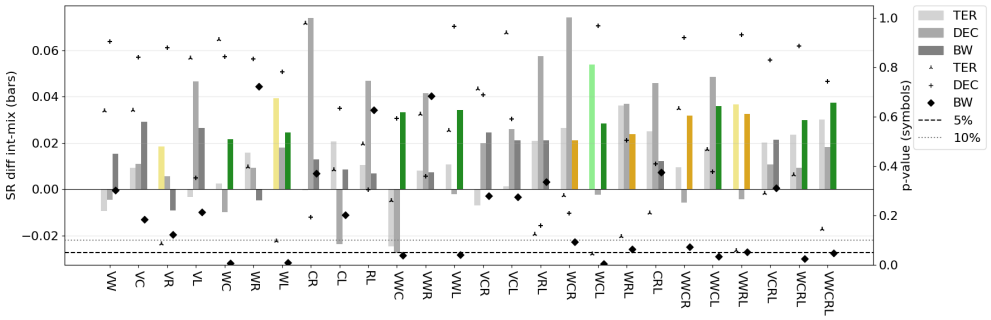


Figure 2: Sharpe ratio differences (integrated minus mixed) with unadjusted p-values. Bars: difference. Symbols: p-value. Dashed: 5%. Dotted: 10%.

Testing 26 hypotheses at once raises the odds of a false rejection. A 5% test repeated 26 times will wrongly reject at least one true null far more than 5% of the time. Figure 3 shows p-values after applying the Romano-Wolf stepdown procedure, which controls the family-wise error rate.

Once we adjust for multiple testing, no combination remains significant at 5%. The few cases that looked promising in Figure 2 fade once the correction is applied. This echoes Leippold et al. [2018]: integrated often wins by a small margin in sample, but the margin is too thin to rule out chance.

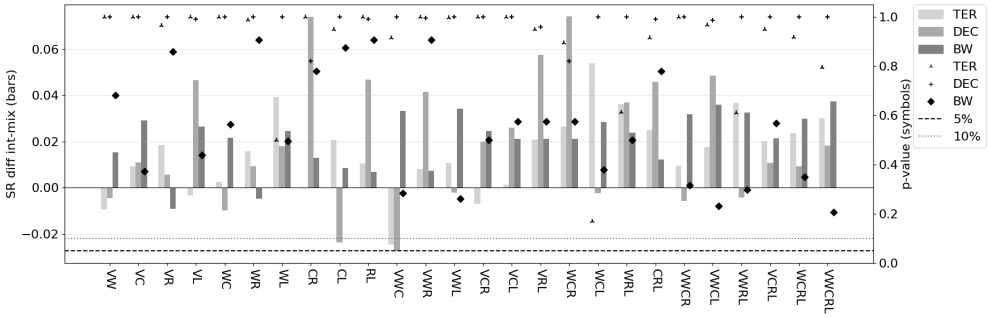
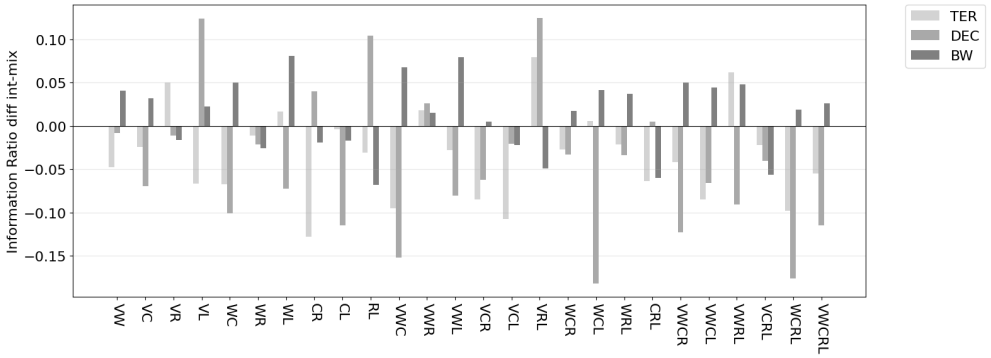


Figure 3: Sharpe ratio differences (integrated minus mixed) with Romano–Wolf adjusted p-values. No combination is significant at 5% after correction.

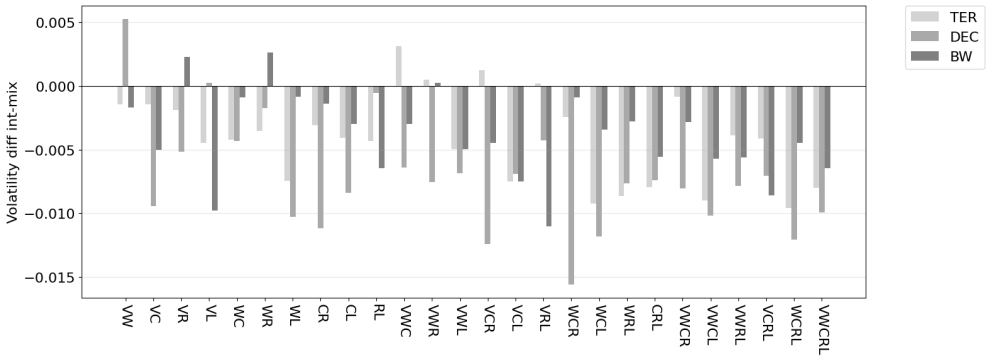
4.1.2 Information Ratio, Variance, and Tracking Error

Figure 4 shows the differences between the integrated and mixed approaches in terms of information ratio, volatility, and tracking error across all portfolios. The

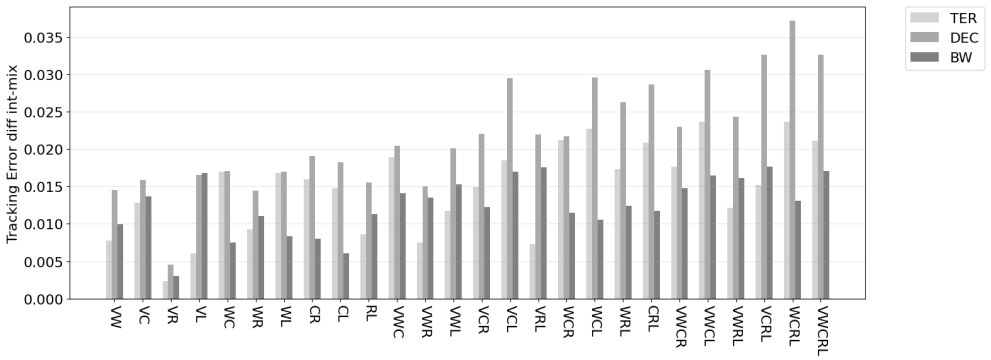
information ratio differences do not display a clear pattern, whereas volatility is consistently lower for the integrated portfolios. At the same time, integrated portfolios exhibit consistently higher tracking error, indicating greater active risk. This implies an unbalanced exposure to risk, which in this setting is associated with larger exposures to factors linked to higher expected returns. As a result, the comparison between the integrated and mixed approaches is not entirely fair. To enable a more comparable assessment, the portfolio construction methods for the mixed approach are adjusted to achieve a similar level of active risk. This is the focus of the next section.



(a) Information Ratio



(b) Annualised Volatility



(c) Tracking Error

Figure 4: Metric differences (integrated minus mixed).

4.2 Adjusted portfolio formation rules

As documented in Section 4.1.2, the integrated and mixed approaches produce portfolios with different risk characteristics even when the same formation rule is applied. The integrated portfolios exhibit lower volatility but higher tracking error, indicating that the two methods take on different levels and types of active risk. This makes the comparison unbalanced. To isolate the effect of signal aggregation from the effect of risk exposure, we adjust the portfolio formation rules so that both approaches operate at comparable risk levels.

4.2.1 Portfolio formation adjustment

For the tercile method, we tighten the selection threshold used by the mixed approach. Rather than selecting the top third for each factor, we select a smaller fraction: 20% for two factors, 12.25% for three, 10% for four, and 8.75% for five. The integrated approach continues to select the top third based on the composite score. The decile method is left unchanged, with both approaches selecting the top 10%. For the Bender–Wang method, we increase the power parameter applied to the mixed portfolios, amplifying the weight differences between high- and low-scoring stocks. For the tracking-error method, integrated portfolios target 2% tracking error while mixed portfolios receive 3–4%, calibrated to produce comparable realised active risk.

4.2.2 Sharpe Ratio

Figure 5 repeats the Sharpe ratio comparison under these adjusted thresholds. The picture changes. Integrated still wins more often than it loses, but the margins shrink and more red tips appear. The tracking-error portfolios show the most mixed results, with neither approach dominating.

Figures 6 and 7 present the statistical tests under the adjusted design. Reading Figure 6, the bars show Sharpe ratio differences (integrated minus mixed) and the symbols show p-values. A symbol below the dashed line indicates significance at 5% before adjusting for multiple testing.

Several patterns stand out. First, the Bender-Wang method produces differences close to zero for nearly every factor combination. This is expected: when all stocks receive some weight, the two approaches converge. Second, the tercile and decile methods show more variation, but the direction is no longer consistent. Some combinations favour integrated, others favour mixed.

Third, the tracking-error method generates the largest swings in both directions. For certain combinations like VCL or VWL, integrated leads by a wide margin; for others like WL or CWCR, mixed comes out ahead. The pattern depends heavily on which factors are included.

Looking at p-values, a handful of combinations cross the 5% threshold in Figure 6. But this is misleading. With 26 tests per method and four methods, we are running

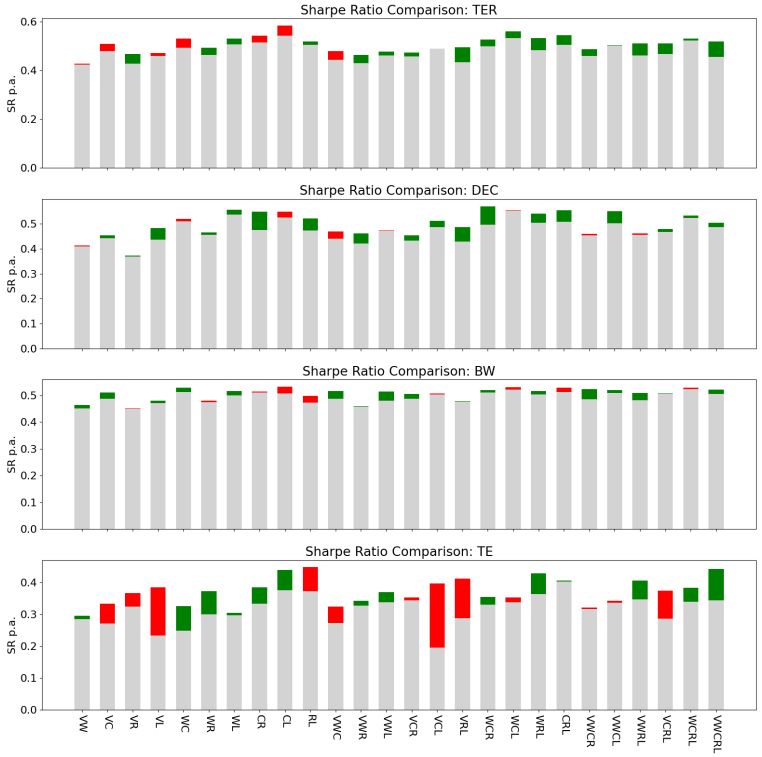


Figure 5: Sharpe ratios after adjusting selection thresholds so that mixed and integrated hold comparable numbers of stocks. Includes tracking-error method.

over 100 comparisons. By chance alone, we expect several to appear significant even if no true difference exists.

Figure 7 applies the Romano–Wolf mutliple testing correction. The adjusted p-values account for the fact that we are testing many hypotheses simultaneously. The dashed line still marks 5%, but now crossing it requires stronger evidence. After correction, no combination remains significant. The symbols that dipped below the threshold in Figure 6 now sit comfortably above it.

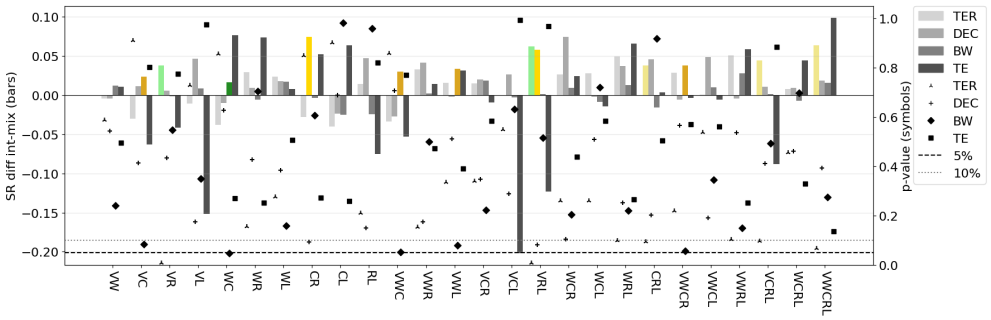


Figure 6: Sharpe ratio differences with unadjusted p-values under adjusted thresholds.

The robustness check reinforces the main finding. When active risk is equalised, the integrated approach loses much of its apparent advantage. What looked like superior signal aggregation turns out to reflect differences in risk exposure rather

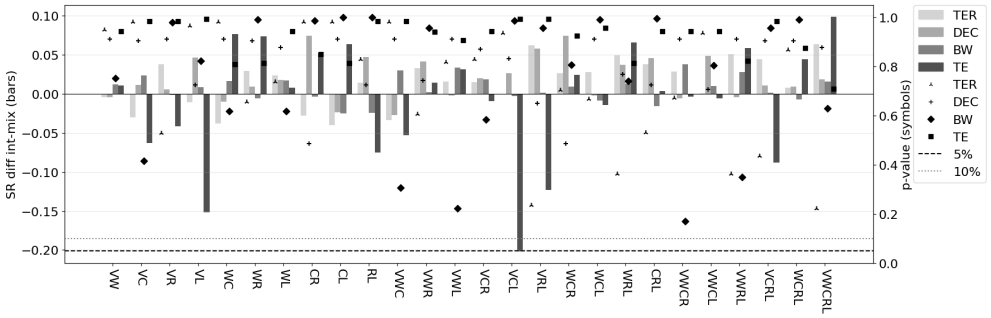
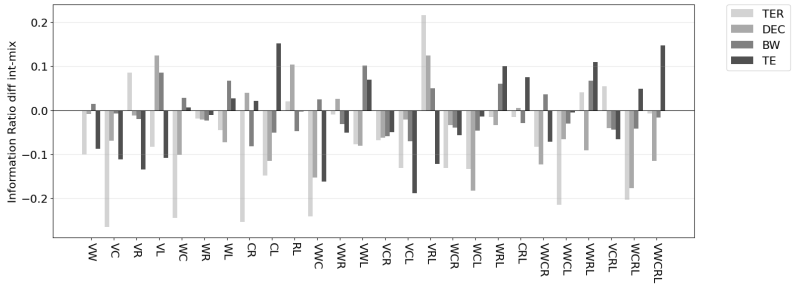


Figure 7: Sharpe ratio differences with Romano–Wolf adjusted p-values under adjusted thresholds. No combination is significant at 5%.

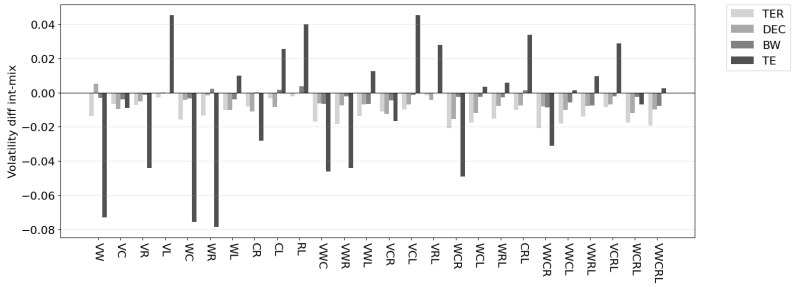
than differences in how the two methods process factor information.

4.2.3 Information Ratio, Variance, and Tracking Error

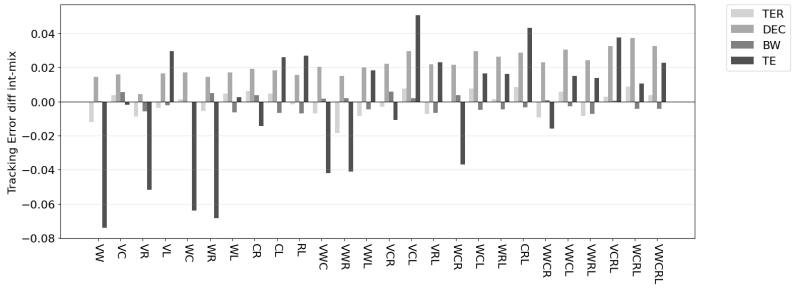
Figure 8 shows the differences between the integrated and mixed approaches in terms of information ratio, volatility, and tracking error across all portfolios after adjusting the mixed portfolios to take on higher active risk. The information ratio now tends to favor the mixed approach, while volatility for the mixed portfolios remains consistently higher than for the integrated portfolios. For the tercile and Bender–Wang portfolios, tracking error is now more comparable between the two approaches. In contrast, portfolios constructed using the tracking-error formation rule exhibit larger absolute differences in both volatility and tracking error. This outcome is expected, as under this formation rule the mixed approach is allowed a higher tracking-error level, which permits greater factor concentration and larger weights in individual stocks with the corresponding characteristics.



(a) Information Ratio



(b) Annualised Volatility



(c) Tracking Error

Figure 8: Metric differences (integrated minus mixed) for after adjusting for comparable active risk.

4.3 Discussion And Conclusion

The empirical evidence leads to a simple conclusion: there is no statistically reliable difference between the mixed and integrated approaches to multi-factor investing. This holds across all 26 factor combinations, all four portfolio construction methods, and both the naïve and concentration-adjusted implementations.

This does not mean the two approaches behave identically. The integrated method consistently produces portfolios with lower volatility. It filters out stocks that rank well on one factor but poorly on others, and what survives tends to be less extreme. In raw Sharpe ratio terms, this volatility reduction often translates into a small edge for integrated. But the edge is not large enough to survive statistical scrutiny. Once we account for the noise in monthly returns and the number of comparisons we are making, the differences vanish.

The choice between mixed and integrated is often framed as a substantive investment decision, with proponents on each side citing theoretical arguments and backtested evidence. Our results suggest the debate may be overblown. Both approaches harvest factor premiums in similar ways, and the performance gap between them is economically small and statistically indistinguishable from zero. An investor choosing between the two should focus on other considerations: implementation costs, turnover, capacity constraints, or simply which approach is easier to explain to clients.

The finding also speaks to a broader issue in quantitative finance. Many strategy comparisons rely on point estimates of Sharpe ratios without asking whether the observed differences could have arisen by chance. A 0.05 difference in annual Sharpe ratio sounds meaningful, but with 60 years of monthly data and substantial return volatility, such differences are well within the range of sampling error. The Romano-Wolf adjustment makes this explicit: when we test many hypotheses at once, we need stronger evidence to declare a winner.

Finally, our results confirm the central finding of Leippold et al. [2018]. Using a different sample period and a slightly different set of implementation choices, we reach the same conclusion they did: the integrated approach does not reliably outperform the mixed approach. If anything, the extended sample through 2024 strengthens the case for agnosticism. Factor investing works, but the details of how you combine factors matter less than you might think.

Author Contributions

All authors jointly contributed to the research design. **Adan Fhima** and **Chedi Mnif** led the methodological analysis and development of the theoretical framework for robust inference, including the bootstrap procedures and multiple testing corrections of Romano et al. [2016]. **Christopher Dybdahl** built the computational pipeline, coding the portfolio construction methods, factor scoring, and statistical tests. All three authors contributed equally to the writing of this report.

References

Markus Leippold, Yunzhi Su, and Chen Wang. Mixed vs. integrated factor portfolios. **Journal of Financial Economics**, 130(2):484–508, 2018. pages 2, 3, 4, 7, 8, 9, 10, 12, 18

Olivier Ledoit and Michael Wolf. Robust performance hypothesis testing with the sharpe ratio. **Journal of Empirical Finance**, 15(5):850–859, 2008. pages 2, 4, 5, 6, 11

Olivier Ledoit and Michael Wolf. Robust performance hypothesis testing with the variance. **Wilmott**, 2011(56):86–89, 2011. pages 2, 5, 6

Joseph Romano, Azeem Shaikh, and Michael Wolf. Efficient computation of adjusted p-values for resampling-based stepdown procedures. **Econometrica**, 84(6):2053–2081, 2016. pages 2, 4, 5, 7, 18

ETFGI. Etfgi reports assets invested in smart beta etfs reach $1.56 trillion, 2024. Industry Report. pages 3

Shaun Fitzgibbons, Jeffrey Friedman, Lukasz Pomorski, and Ludovic Serban. Long-only style investing: Don't just mix, integrate. **The Journal of Portfolio Management**, 42(5):31–46, 2016. pages 3, 10

Roger Clarke, Harindra de Silva, and Steven Thorley. The factor zoo: A statistical examination. **Financial Analysts Journal**, 72(4):32–49, 2016. pages 3

Campbell R Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. **The Review of Financial Studies**, 29(1):5–68, 2016. pages 3, 7

David Bailey, Jonathan Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism. **Notices of the American Mathematical Society**, 61(5):458–471, 2014. pages 3, 7

Joseph Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. **Econometrica**, 73(4):1237–1282, 2005. pages 3

David Blitz and Pim van Vliet. The volatility effect: Lower risk without lower return. **Journal of Portfolio Management**, 34(1):102–113, 2007. pages 4

Donald W. K. Andrews and John C. Monahan. Improved HAC estimation of the long-run variance. **Econometrica**, 60(4):953–966, 1992. pages 6

Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. **The Journal of Finance**, 47(2):427–465, 1992. pages 8, 10

Eugene F Fama and Kenneth R French. A five-factor asset pricing model. **Journal of Financial Economics**, 116(1):1–22, 2015. pages 9

Jennifer Bender and Taie Wang. Can the whole be more than the sum of the parts? bottom-up versus top-down multifactor portfolio construction. **Journal of Portfolio Management**, 42(5):39–50, 2016. pages 10