Analyzing the Performance of Multi-Factor Investment Strategies under Multiple Testing Framework

Yu-Chin Hsu
Institute of Economics
Academia Sinica

Hsiou-Wei Lin
Department of International Business
National Taiwan University

Kendro Vincent
Department of International Business
National Taiwan University

Correspondence to Kendro Vincent, Department of International Business, National Taiwan University; E-mail: vincent.kendro@gmail.com.

Analyzing the Performance of Multi-Factor Investment Strategies under Multiple Testing Framework

Abstract

Evaluating portfolios based on numerous combinations of factors using individual backtesting method could suffer serious data mining bias and lead to spurious significant findings. Accordingly, we employ a multiple testing method to examine the significance of Sharpe ratios of multi-factor portfolios. Our empirical results show that even after adjusting for data mining bias, the stock picking strategies with certain combined firm characteristics could obtain significantly better risk-scaled returns than both value-weighted index and small-cap value portfolio. The superior performance of multi-factor portfolios is more stable than single factor portfolios over different subsamples. Moreover, the outperforming multi-factor strategies are robust to alternative definitions of factors.

Keywords: Data mining bias, multi-factor investment strategy, multiple hypotheses testing, passive index investing, smart beta.

JEL classifications: G11, G17.

Introduction

Combining multiple factors to build winning stock picking strategy is an appealing idea for a growing and large number of investors. Unfortunately, since there are numerous possible combinations of factors, the backtesting of multi-factor strategies based on individual hypothesis testing framework could lead to serious data mining bias. That is, if the researcher uses 5% significance level to test each individual hypotheses with the same historical datasets, she could eventually find a spuriously outperforming strategy after a large number of trials.

To address the data mining bias problem, we apply bootstrap-based multiple testing method to evaluate the performance of multi-factor investment strategies, which have been quite popular in recent years. The ETFs under the name of smart beta or multi-factor have attracted over $60 billion fund inflows each year since 2013 and account for one-fifth of the $1.7 trillion US ETF total assets in 2015 (Wigglesworth 2016). Kahn and Lemmon (2016) even state that the smart beta product innovation has posed a threat to the traditional active fund industry. While the current majority of smart beta ETFs adopt an uncomplicated rule based on the single factor such as value strategy, there has been a trend for ETF providers to launch products that combine multiple factors. This type of ETFs has been dubbed smart beta 2.0, smarter beta, or multi-factor funds (Authers 2015; Noblett 2015; Wigglesworth 2016).

The concern for data mining bias is not new. McQueen and Thorley (1999) discuss the potential data mining pitfalls in testing the portfolio performance of quantitative stock screening. Bailey and de Prado (2014) and Harvey and Liu (2014) provide the statistical framework to adjust the Sharpe ratio for evaluating trading strategies in general. Novy-Marx (2016) shows that the *t*-statistics of mean returns could be severely inflated when the portfolio is constructed by combining the best strategies.

The contribution of this study is twofold. First, we apply the data-driven approach by Hsu et al. (2014) to empirically examine the performance of the multi-factor strategies. The aforementioned studies focus on adjusting the backtesting to mitigate data mining bias but do not conduct an empirical investigation on multi-factor strategies. In contrast, this study answers the question of whether there exists superior multi-factor strategy after adjusting for data mining bias. Furthermore, because the multiple testing methods that are based on independence or arbitrary dependence assumption could be much more conservative, they could have much less power to identify significantly superior strategies. The advantage of data-driven methodology is that the information about dependence structure among the portfolios performance is taken into account directly from the data and hence is more powerful.

Second, we show that there are benefits by combining a few factors into one while constructing portfolio strategy. For instances, the order of outperformance in the single-factor portfolios universe is sensitive to the definition of factors, whereas the outperforming multi-factor strategies are more robust. A recurring theme is the portfolio of firms with combined characteristics of small-cap, value, high momentum, and low volatility; the portfolio's outperformance is consistently significant regardless how we measure valuation ratio, momentum, or volatility. The in-sample and out-of-sample analysis also show that multi-factor strategies have more stable outperformance than the single-factor portfolios. This is because the performance of multi-factor strategies does not rely on one particular factor exposure, instead the constituents of the portfolios are firms with diversified characteristics to earn higher risk-adjusted returns. This finding is also consistent with the suggestion by Cliff Asness, co-founder of AQR Capital Management, who advocates diversification and keeping all factors "on" most of the time (Kim, 2016).

By using the sample from 1968 to 2015, we identify quite a few multi-factor strategies that significantly outperform market portfolio in terms of Sharpe ratio under multiple hypotheses testing. The critical value for the annualized Sharpe ratio, which is free of data mining bias, is approximately 0.6. We consider 8 different classes of factors to combine and each portfolio is rebalanced annually. Our results show that stock selection strategies based on combining 4 to 6 factors generate significant superior returns and the results are robust against different choices of portfolio weights or definitions of factors. The annualized Sharpe ratio could be as high as 1 in the in-sample analysis.

Furthermore, we examine the outperformance of the multi-factor portfolios against small-cap value portfolio. Arnott et al. (2013) suggest that many multi-factor strategies or smarter beta ETFs resemble small-cap value portfolio, thus they simply provide product varieties without additional economic value. The critical value for the annualized Sharpe ratio increases to around 0.8 if we change the benchmark to small-cap value portfolio. As a result, the number of outperforming multi-factor portfolios is significantly reduced. Nevertheless, we still find that there exists ample opportunity for investors to improve their portfolio selection through multi-factor strategies.

Backtesting without Data Mining Bias

We use the difference between Sharpe ratio of the i-th portfolio, SR^i, and that of the benchmark portfolio, SR^b, as the portfolio performance evaluation metric. Our null hypotheses of interests are

$$\mathbf{H}_\mathbf{0}^i\colon \mathrm{SR}^i - \mathrm{SR}^b \leq 0, \qquad i = 1, \dots, \mathrm{M}, \tag{2}$$

where M is the number of portfolios considered and SR is the annualized Sharpe ratio calculated as $\sqrt{12}$ multiply with the mean excess returns and then divide by the standard deviation of the excess returns. Our objective is to find a common threshold or critical value to decide which null hypotheses to reject under multiple testing framework. In other words, we thereby determine how large the Sharpe ratio difference should be for a collection of multi-factor portfolios to significantly beat the benchmark portfolio without data mining bias.

The critical value for a test statistic in individual hypothesis testing is estimated so that the probability of committing Type I Error is bounded below a certain threshold which depends on the desired significance level. However, in multiple hypotheses testing, we need to consider a different notion of error rate before we develop the testing procedure. Unlike individual hypothesis testing, there are various types of error rates in the multiple testing literature. In this study, we use False Discovery Proportion (FDP), which is defined as the ratio between the number of false rejections (Type I Errors) and the total number of rejections. If there is no rejection, then FDP is defined as 0. We say that a statistical testing method controls the FDP if the procedure could assure

$$\mathrm{P}\{\mathrm{FDP} > \xi\} \leq \alpha, \tag{1}$$

where ξ and α are the user-specified inputs.

To control the FDP, we use the FDP-SPA procedure outlined in Hsu et al. (2014). Intuitively, the multiple testing framework proceeds as follows. First, we calculate the performance measures for portfolios $i = 1, \dots, \mathrm{M}$. Starting from $k = 1$, we follow Step-SPA(k) algorithm also in Hsu et al. (2014), which controls k-FWER at α level, where k-FWER is the probability of falsely rejecting at least k true null hypotheses. We then reject the portfolios with performance metrics being greater than the critical value estimated by Step-SPA(1). If there is no rejection, i.e. none of the portfolio has superior performance, then we stop. Otherwise, we apply Step-SPA(k+1) until $k/(N_k + 1) > \xi$, where N_k is the number of rejection at stage k. The final critical value is the threshold for the portfolio performance measures that would asymptotically control the FDP below α. The exact algorithms of the Step-SPA(k) and FDP-SPA are given in the Appendix.

In contrast to the procedure suggested in Harvey and Liu (2014) and Harvey et al. (2016), our method takes into account the dependence structure among the portfolio performances directly from the data. Specifically, the methodology involves a bootstrap estimation of the probability distribution in each step of the stepwise procedure. Moreover, Romano and Wolf (2007) suggest that if the test statistics are correlated, then the distribution of FDP could be highly skewed. This would cause the False Discovery Rate (FDR)[1] controlling method, which concerns only about the mean of FDP, be a less appropriate choice of error rate. Another advantage of the Hsu et al. (2014)'s procedure is that it could minimize the impact of irrelevant underperforming multi-factor strategies, because it is shown in Hansen (2005) that the statistical power of multiple hypothesis testing could be substantially reduced if too many irrelevant inferior models are included.

Data and Construction of Multi-Factor Strategies

We obtain the accounting and monthly stock returns data of US firms over 1968-2015 from Compustat and CRSP databases, respectively. We follow Beaver et al. (2007) to adjust for delisting bias in CRSP stock returns data. To ensure that our sample on US stocks could serve as meaningful trading purposes, we exclude firms with negative book value of

[1] FDR is defined as conditional expected value of FDP on total number of rejection is greater than zero.

equity and stock price below $1. The risk-free rate is the one-month Treasury bill rate obtained from Professor Kenneth French's website.

A preliminary step in the multiple testing is to choose a set of factors and then construct the universe of multi-factor portfolios. Over the past two decades, the empirical finance literature has discovered several factors affecting stock returns (see, e.g., Hou et al. 2015). We select 8 styles of investment strategy and pick one representative variable for each style as the factor to construct the multi-factor portfolios.

Table 1 gives the definition of the factors we adopt. The multi-factor ETF usually employ four categories of factors. For example, JP Morgan Diversified Return US Equity ETF and Global X Scientific Beta U.S. ETF use value, size, momentum, and volatility as their underlying factors, and iShares FactorSelect MSCI US ETF combines value, size, momentum, and quality. We choose to adopt a broader classification than the current practices of multi-factor ETFs. Other categories or factors that are not covered by major multi-factor ETF providers have been documented in the academic literature to be influential stock return determinants. By including a richer set of investment styles, we can also test whether the performance of multi-factor strategy could be improved by combining more factors.

Our multi-factor portfolios are constructed as follows. Each factor is ranked according to its value in ascending order at the end of April every year. We choose April as the cut-off for portfolio formation because we retain only firms with December fiscal year end and assume the accounting information is available with a four-month lag. After the factors are sorted independently, we assign a variable $\mathrm{r}(x^l)$ equals the value from 1 to N (the last number of observations) for each factor *x* based on its rank. The superscript *l* indicates that firm with the lower value of factor receives lower score. If the ranks are tied, then we will assign the same score. We also define another high-to-low ranked variable $\mathrm{r}(x^h)$ where its value equals 1 to N for each decreasingly ranked factor, i.e. firms with greater value of factor receives lower score. We include the "upside-down" version since we choose not to take any prior knowledge on which order should the factors be ranked. Moreover, Arnott et al. (2013) find that some stock selection strategies based on sensible investment belief, such as picking low PE ratio stocks strategy, could have equal performance when the factors are sorted in a reverse direction.

We then transform r(x) into z-score, z(x), by subtracting its mean and dividing by its standard deviation within the same year. After calculating the z-scores, we add them together with various combinations. The multi-factor portfolios are constructed by longing the stocks with combined z-scores being within the bottom decile. For instance, small-cap value momentum strategy forms the portfolio of stocks with combined z-scores, $\mathrm{z}(mve^{l}) + \mathrm{z}(bm^{h}) + \mathrm{z}(mom12^{h})$, being below 10% quantile at the portfolio formation time.

The portfolio weights for each stock are calculated as follows. Let $z_1, \dots, z_S$ denote the combined z-score of the selected *S* stocks. We rank and standardize the values of $z_1, \dots, z_S$ once again and denote the new score with $v_1, \dots, v_S$. The portfolio weight of stock *i* at the portfolio formation time is defined as

$$w_i = \frac{1 - \Phi(v_i)}{\sum_{j=1}^{S} 1 - \Phi(v_j)}, \qquad i = 1, \dots, S, \tag{3}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function. We construct the portfolio weight this way so that the portfolio exposure to multi-factor score is greater. This is also in line with Arnott et al. (2005).

Main Results

We begin by analyzing the results of single-factor portfolios. The value-weighted market portfolio (*mkt*) is used as our benchmark. Figure 1 plots both mean and standard deviation of the portfolio excess returns. The factor portfolio returns in our sample exhibit the same pattern as those predicted in the literature. For instance, the portfolio of firms with *bm* in the high decile outperforms the ones in the low decile. The small-cap portfolio (mve^{l}) has greater mean returns than large-cap portfolio (mve^{h}). The low volatility portfolio ($tvol^{l}$) has much smaller standard deviation than high volatility portfolio ($tvol^{h}$), despite that their mean returns are approximately the same. Excluding the portfolio $tvol^{l}$, the market portfolio has the lowest standard deviation. Since we construct the portfolio so that its exposure to the underlying factor is higher, it is not surprising that the portfolio mve^{h} almost has the same mean and standard deviation as market portfolio.

Table 2 shows the critical values for (annualized) Sharpe ratio difference based on the multiple hypotheses testing. We control the FDP to be below 10% with a 95% confidence level, i.e. we set ξ and α in Equation (1) equal 0.1 and 0.05, respectively. One notable result is that the critical values are roughly the same for the all types of universes. The critical value is approximately 0.245, except for the universes of two-factor and eight-factor portfolios which have slightly greater critical values. The result means that to guard against data mining bias, we need to require the multi-factor portfolio's Sharpe ratio to exceed the market portfolio's by 0.245 to claim a significant outperformance. The market portfolio has Sharpe ratio of 0.366 in our sample, so the critical value can also be translated as the hurdle rate for Sharpe ratio is approximately 0.611.

Intuitively, when there is a larger universe of portfolio strategies to search for the significant outperformances, one might expect that the critical value to be greater to guard against data mining. This is in contrast to our result of "flat" critical values across different types of combinations. This is generally true if the performances of the portfolio strategies are independent. However, due to the positive correlation among the factors, the critical value does not need to be greater to alleviate the data mining bias. Another reason for why the critical value is not increasing with the size of portfolio universe is that a lot of the additional strategies which span the larger universe have negative returns, therefore their contribution to the distribution becomes irrelevant in the limit[2].

The numbers of outperforming strategies presented in Table 2 show that there are various multi-factor investment strategies with statistically significant Sharpe ratio differences. In each of the five-factor or six-factor strategies alone, there are almost 200 combinations that could outperform the market portfolio. In Table 3 we report the best 5 outperforming schemes for each multi-factor strategy, the number in parenthesis is the respective Sharpe ratio difference.

The low volatility portfolio is the only single-factor strategy with Sharpe ratio difference being greater than the corresponding critical value. The outperformance of low volatility strategy is also apparent from Figure 1, it is the only factor portfolio with greater mean

[2] This is one of the properties in Hsu et al. (2014)'s FDP-SPA. For detailed descriptions, see Appendix.

excess returns and lower standard deviation than the market portfolio. The other outperforming factor portfolios mostly lie on the upper right to the market portfolio, i.e. they beat market portfolio in terms of mean excess returns, but their standard deviations are also higher than the market portfolio's. Note that the Sharpe ratio of small-cap portfolio or high book-to-market portfolio is also considerably greater than the market portfolio's but is not large enough to surpass the critical value. This shows that after adjusting for data mining bias, the returns of most single-factor portfolios are statistically insignificant.

For the two-factor strategies, there are 11 out of 112 portfolios with the values of Sharpe ratio difference above the critical value (0.255). Table 3 shows that all of the best 5 two-factor portfolios involve low volatility factor in their sorting strategy. In addition to outperforming the benchmark with a significant difference in Sharpe ratios, the improvements over low volatility portfolio are quite modest. The Sharpe ratio differences of well-known two-factor strategies such as small-cap value (bm^h/mve^l) and momentum value ($bm^h/mom12^h$) are much less than portfolios which include low volatility factor, but they are still greater than the critical value (0.168 and 0.341, respectively).

We find that by combining three or four factors, we could increase the Sharpe ratio difference over low volatility portfolio's by a much higher margin. All of the best 5 outperforming strategies have at least 0.5 Sharpe ratio difference. It is also noteworthy that the best strategy in four-factor universe resembles certain popular ETF products mentioned in the introduction. The result shows that portfolio based on combined ranking of small-cap, value, momentum, and low volatility ($bm^h/mve^l/mom12^h/tvol^l$) delivers Sharpe ratio of 0.579 in excess of the market portfolio Sharpe ratio. If we drop market capitalization from the combination, the Sharpe ratio difference becomes 0.512, which is still statistically significant.

To compare the portfolio performances across different number of factors, we plot the Sharpe ratio differences against the rankings within each type of multi-factor strategy. Figure 2 shows that there is a modest gain in Sharpe ratio difference by adding one more factor to form four-factor strategy. The top 3 outperforming six-factor portfolios and the best seven-factor portfolio have the highest Sharpe ratio among all multi-factor strategies. The remaining six-factor portfolios fail to perform as well as the five-factor portfolio.

Moreover, most of the seven-factor portfolios are inferior to five-factor or four-factor portfolios. This pattern is also similar to the eight-factor strategies. The best of eight-factor portfolios outperforms the market portfolio by the similar magnitude to the best of four-factor portfolios. However, the performances of the remaining eight-factor portfolios deteriorate sharply. Some factors in the eight-factor strategies are combined using the upside-down version, for instance, Table 3 shows that its third best strategy includes low momentum factor ($mom12^l$) and the fourth best strategy applies low profitability (roe^l) instead of high profitability factor (roe^h).

Out-of-Sample Performance. In Table 4, we compare the in-sample and out-of-sample portfolio risk-scaled returns. We choose May 2010 as the cutoff to split the sample into the in-sample and out-of-sample period. McLean and Pontiff (2016) suggest that many of the factors premium or the outperformance over market portfolio disappear in the sample period after their academic publications. All of the factors we use in this study are published before 2010, hence the cutoff in May 2010 should be a reasonable choice to evaluate out-of-sample performances. For the sake of brevity, we show the results of single- to five-factor portfolios, the rest of the multi-factor portfolios unreported here paint the similar picture.

Panel A of Table 4 reports the top 3 outperforming strategies in each multi-factor portfolio universes for the in-sample and out-of-sample period. For each strategy, we show the in-sample and out-of-sample Sharpe ratio differences in the parentheses, respectively. The in-sample critical values for Sharpe ratio differences, shown in Panel B, are approximately 0.26 which is slightly higher than the estimates based on the full sample. However, we note that the market portfolio Sharpe ratio is 0.304 based on the sample prior to May 2010, so the data-mining bias free threshold for the Sharpe ratio is roughly 0.57, which is less than the one estimated with full sample data.

With the exception of low volatility portfolio, all of the single-factor portfolios perform worse than market portfolio in the out-of-sample period. The small-cap portfolio and value portfolio are the second and third best portfolios in-sample but have negative Sharpe ratio differences and are not ranked within the top 3 portfolios in the out-of-sample period. The results also show that large-cap portfolio outperforms small-cap portfolio after May 2010; however, both of them underperform market portfolio.

The out-of-sample performances of multi-factor portfolios are relatively more stable than the single-factor portfolios. It appears to be difficult to predict which strategy would deliver the highest Sharpe ratio difference. Nevertheless, the best strategy according to in-sample Sharpe ratio difference is still among the top outperforming strategies in the out-of-sample period. We also note that the Sharpe ratio of market portfolio in the out-of-sample period (0.912) is much greater than the in-sample period (0.304); therefore, even if the multi-factor strategies outperform the market portfolio by lower magnitude after May 2010, the absolute Sharpe ratio of the portfolio could still be greater than the in-sample Sharpe ratio.

Are Multi-Factor Strategies Really Smarter? Many skeptics suggest that multi-factor strategies are merely a marketing gimmick of ETF vendors. Arnott et al. (2013) argue that many multi-factor strategies are actually portfolios with exposure tilted toward small-cap value factors, therefore further adding different factors would not enhance portfolio performance. In Table 5, we report the backtesting results by replacing the benchmark with small-cap value portfolio to investigate if there are still any outperformances in multi-factor portfolios.

In contrast to the results when the market portfolio is used as benchmark, the numbers of significant outperforming strategies are considerably much smaller since the Sharpe ratio threshold now becomes greater. The critical values for Sharpe ratio differences range between 0.286 and 0.301. The Sharpe ratio of small-cap value portfolio in our full sample period is 0.534. This means that the absolute Sharpe ratio of the multi-factor portfolio has to be at least 0.82 to be deemed statistically significant under our multiple hypotheses testing framework. Therefore, there are not many strategies that could beat small-cap value portfolio. This result is as suggested by Arnott et al. (2013). Nevertheless, the evidences in Table 5 suggest that other factors such as low volatility and high momentum help improve the portfolio performances. For example, the Sharpe ratio difference of portfolio $bm^h/mve^l/mom12^h/tvol^l$ and small-cap value portfolio is 0.412, which surpasses the critical value for four-factor portfolio universe.

Robustness Check

In this section, we discuss the sensitivity of the main results with respect to various research design modifications.

Trimming the extreme observations. To examine if the significant outperformance is driven only by a few outliers, we exclude the firms with the extreme value of factors before calculating the combined multi-factor score. For each factor, we trim the 5% extreme observations from the sample. Therefore, if any of the factors is within 5% quantile, then the firms are excluded from the portfolio sorting that year.

The results, not shown here, are fairly similar to the ones in Tables 2 and 3. The critical values of Sharpe ratio differences range between 0.231 and 0.248. The only significant single-factor portfolio is $tvol^l$ with Sharpe ratio 0.391 higher than market portfolio. When we keep the extreme observations, $tvol^l$ has Sharpe ratio difference of 0.381. We also obtain considerably larger number of significant outperforming portfolios compared to the case without trimmed observations. This shows that the superior performance of multi-factor investment strategies is not due to the influence of firms with extreme factors.

Fixed number of stocks portfolio. Since the earlier years have fewer public firms, the selected stocks in the portfolio based on 10% quantile will be less during this period. Instead of choosing the cut-off quantiles, we can fix the number of stocks to include in the portfolio across all years. There are trade-offs in choosing the portfolio size. While lowering the number of stocks to purchase could increase the exposure to the desired factors, it could be less diversified. We try various numbers of stocks: 50, 100, 200, and 300. To save space, we do not report the results here but they are available upon request.

We find that the critical values for portfolios with 50 stocks are around 0.29, which is greater than the critical values shown in Table 2. As a result, the number of significant outperforming portfolios is smaller. Nonetheless, there are still many portfolios with significant Sharpe ratio differences. As we increase the number of selected stocks, the critical value becomes lower and the number of significant portfolios increases. Of note, the top outperforming multi-factor portfolios shown in Table 3, e.g. $bm^h/mve^l/mom12^h/tvol^l$, remain significant regardless the number of stocks included.

Alternative definition of factors. We use a different set of proxy variables to check the robustness of particular investment styles to the definition of their factors. Hsu et al. (2015) also recommended to slightly perturb the definition of the factors while evaluating smart beta strategies. Panel A of Table 6 shows the definition of the new set of factors.

In Panel B and Panel C of Table 6, we conduct the same analysis with this alternative definition of the factors. For results with the market portfolio as the benchmark, the critical values range between 0.23 and 0.24, which is slightly lower than the baseline results in Table 2. Moreover, there are a large number of significant portfolios. The four-factor strategy alone has 245 portfolios that could beat the market portfolio significantly and there are two single-factor portfolios with significant Sharpe ratio differences, momentum ($wh52^h$) and value (ey^h). The low volatility factor portfolio becomes insignificant when we use beta as the proxy variable despite that it still has greater Sharpe ratio than the market portfolio. If we change the benchmark to small-cap value portfolio, then there are only ten portfolios with significant superior performances. However, we still find that the portfolio based on small-cap, value, momentum, and low volatility strategy yields highly significant Sharpe ratio.

The overall results suggest that which single-factor portfolio has the highest Sharpe ratio may depend on the definition of the factor, but as we combine the factors, the performance of multi-factor portfolio becomes more robust to the factor definitions.

Value-weighted strategies. Table 7 presents the results for the multiple testing of portfolio performance when market capitalization is used as the portfolio weight. The estimated critical values for multi-factor portfolio Sharpe ratio difference turn out to be greater than the baseline results, while the single-factor portfolio has smaller estimated critical value. All of the portfolio performances deteriorate after we switch the portfolio weight to market capitalization. None of the single-factor portfolios has a significant Sharpe ratio difference. The results support the fact shown in Arnott et al. (2005) that score-weighted strategy is the better portfolio weighting scheme.

Conclusion

This study aims at recent multi-factor ETFs, which provide investors the opportunity to increase their portfolio exposure beyond both size and value factors. The multiple testing framework allows us to analyze the performance of portfolios which are constructed from numerous factors combinations without data mining bias. The results shown in this study suggest that investors may achieve higher Sharpe ratio through multi-factor portfolio strategies. We consider two benchmark portfolios: value-weighted market portfolio and small-cap value portfolio. The Sharpe ratios of the best performing multi-factor portfolios exceed the benchmarks' Sharpe ratio by a modest magnitude above the estimated critical value.

We find that a strategy of purposefully biasing the portfolio weights toward greater exposure of the multi-factor scores generates better portfolio performance than the market capitalization weighted strategy. Our results also suggest that the performance of multi-factor portfolios remains relatively stable than the single-factor portfolio performance when we alter the definitions of factors. Moreover, while it is not possible to predict which portfolio would have the highest Sharpe ratio ex-ante, the outperforming multi-factor portfolios could still consistently beat the benchmark in both in-sample and out-of-sample periods.

Appendix

This appendix presents the FDP-SPA controlling procedure to test the multiple inequalities

$$\mathbf{H}_{\mathbf{0}}^{i}\colon \mathrm{SR}^{i} - \mathrm{SR}^{b} \leq 0, \qquad i = 1, \dots, M.$$

The FDP-SPA procedure is based on applying Step-SPA(k) algorithm recursively. Let $max(A, k)$ denotes the k-th largest value of vector A and $1(\cdot)$ denotes the indicator function. Furthermore, X_i and Y denote the vector of excess returns of portfolio i and benchmark portfolio, respectively, and let T be the sample size of the excess returns. The procedure ensures that $\mathrm{P}\{\mathrm{FDP} > \xi\} \leq \alpha$ asymptotically, which means that the probability of false discovery proportion exceeding ξ is bounded below α when T is large enough. In our empirical analysis, we use $\xi = 0.1$ and $\alpha = 0.05$. The output of the FDP controlling procedure is a common cutoff point or critical value that determines which portfolios have significantly greater Sharpe ratio than the benchmark portfolio does.

Step-SPA(k) algorithm with level α

---

procedure stepSPA($\{X_1, \dots, X_M, Y\}, \alpha, k$)
Input data: $\{X_1, \dots, X_M, Y\}$ ◆ Data of portfolio and benchmark excess returns
Input parameter: α, k ◆ FWER(k)'s parameters
create vector STAT of size M
for $i \in \{1, \dots, M\}$ **do**
    $\Delta_i = \mathrm{SR}^i - \mathrm{SR}^b$
    $\Delta_i^- = \Delta_i \times 1(\sqrt{T}\Delta_i \leq -\sigma_i\sqrt{2\log\log T})$ ◆ σ_i is the standard error of Δ_i
    STAT$[i]$ = Δ_i
end for
create matrix X with row size M and column size B
for $s \in \{1, \dots, B\}$ **do**
    generate bootstrap sample $\{X_1^s, \dots, X_M^s, Y^s\}$
    for $i \in \{1, \dots, M\}$ **do**
        X$[i, s]$ = $\Delta_i^s - \Delta_i + \Delta_i^-$ ◆ Δ_i^s is the SR difference using bootstrap sample s
    end for
end for
create sort_index which order the vector STAT from high to low
SORTED_X = X[sort_index, :] ◆ re-order rows of X according to the sort_index

NUM_REJECT = 0
NUM_REJECT = -1
create vector KMAX of size B
while NUM_REJECT > NUM_REJECT1 **do** ◆ The procedure will stop when there is no further rejections
    NUM_REJECT1 = NUM_REJECT
    if NUM_REJECT < k **then do**
        for $s \in \{1, \dots, B\}$ **do**
            KMAX[s] = max(SORTED_X[: , s], k)
        end for
    else do
        for $s \in \{1, \dots, B\}$ **do**
            KMAX[s] = max(SORTED_X[(NUM_REJECT$-k+2$):M, s], k)
        end for
    end if
    q = max(KMAX, $round(\alpha \times B)$)
    if q < 0 **then** q = 0 **end if**
    CRITCAL_VALUE = q
    NUM_REJECT = sum(1(STAT > CRITICAL_VALUE))
end while
Output: CRITICAL_VALUE
end procedure

FDP-SPA with α and ξ

procedure FDP_SPA($\{X_1, \dots, X_M, Y\}, \alpha, \xi$)
Input data: $\{X_1, \dots, X_M, Y\}$ ◆ Data of portfolio and benchmark excess returns
Input parameter: α, ξ ◆ FDP-SPA's parameters
create vector STAT of size M
for $i \in \{1, \dots, M\}$ **do**
    STAT[i] = $\mathrm{SR}^i - \mathrm{SR}^b$
end do
SPA_k = 1
CRITICAL_VALUE = stepSPA($\{X_1, \dots, X_M, Y\}, \alpha$, SPA_k)
NUM_REJECT = sum(1(STAT > CRITICAL_VALUE))
while NUM_REJECT < SPA_k / $\xi - 1$ **do**

SPA_k = SPA_k + 1

CRITICAL_VALUE = stepSPA($\{X_1, \ldots, X_M, Y\}, \alpha,$ SPA_k)

NUM_REJECT = sum(1(STAT > CRITICAL_VALUE))

end while

Output: CRITICAL_VALUE

end procedure

Reference

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

Arnott, R. D., Hsu, J., and Moore, P. (2005). "Fundamental indexation." *Financial Analysts Journal*, **61**, 83-99.

Arnott, R. D., Hsu, J., Kalesnik, V., and Tindall, P. (2013). "The surprising alpha from Malkiel's monkey and upside-down strategies." *Journal of Portfolio Management*, **39**, 91-105.

Authers, J. "Why multi-factor funds are smarter beta." *Financial Times*, May 13, 2015. http://on.ft.com/1e0F6wa (accessed March 21, 2016).

Bailey, D. H. and de Prado, M. L. (2014). "The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality." *Journal of Portfolio Management*, **40**, 94-107.

Banz, R. W. (1981). "The relationship between return and market value of common stocks." *Journal of Financial Economics*, **9**, 3-18.

Beaver, W., McNichols, M., and Price, R. (2007). "Delisting returns and their effect on accounting-based market anomalies." *Journal of Accounting and Economics*, **43**, 341-368.

DeBondt, W. F. M. and Thaler, R. (1985). "Does the stock market overreact?" *Journal of Finance*, **40**, 557-581.

Cooper, M. J., Gulen, H. and Schill, M. J. (2008). "Asset growth and the cross-section of stock returns." *Journal of Finance,* **63**, 1609-1651.

Fama, E. F. and French, K. R. (2006). "Profitability, investment and average returns." *Journal of Financial Economics*, **82**, 491-518.

Hansen, P. R. (2005). "A test for superior predictive ability." *Journal of Business & Economic Statistics*, **23**, 365-380.

Harvey, C. R. and Liu, Y. (2014). "Evaluating trading strategies." *Journal of Portfolio Management*, **40**, 108-118.

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

Hou, K., Xue, C., and Zhang, L. (2015). "Digesting anomalies: An investment approach." *Review of Financial Studies*, **28**, 650-705.

Hsu, J., Kalesnik, V., and Viswanathan, V. (2015). "A framework for assessing factors and implementing smart beta strategies." *Journal of Index Investing*, **6**,89-97.

Hsu, Y.-C., Kuan, C.-M., and Yen, M.-F. (2014). "A generalized stepwise procedure with improved power for multiple inequalities testing." *Journal of Financial Econometrics*, **12**, 730-755.

Jeegadessh, N. (1990). "Evidence of predictable behavior of security returns." *Journal of Finance*, **45**, 881-898.

Kahn, R. N. and Lemmon, M. (2016). "The asset manager's dilemma: How smart beta is disrupting the investment management industry." *Financial Analysts Journal*, **72**, 15-20.

Kim, C. (2016). Two ways to beat the S&P. *Barron's*, **96**, 25.

McLean, R. D. and Pontiff, J. (2016). "Does academic research destroy stock return predictability?" *Journal of Finance*, **71**, 5-32.

McQueen, G. and Thorley, S. (1999). "Mining fool's gold." *Financial Analysts Journal*, **55**, 61-72.

Noblett, J. "Smart beta's potential starting to overwhelm ETF providers." *Financial Times*, December 20, 2015. http://on.ft.com/1QTyw9Y (accessed March 21, 2016).

Novy-Marx, R. (2016). "Backtesting strategies based on multiple signals." Working paper.

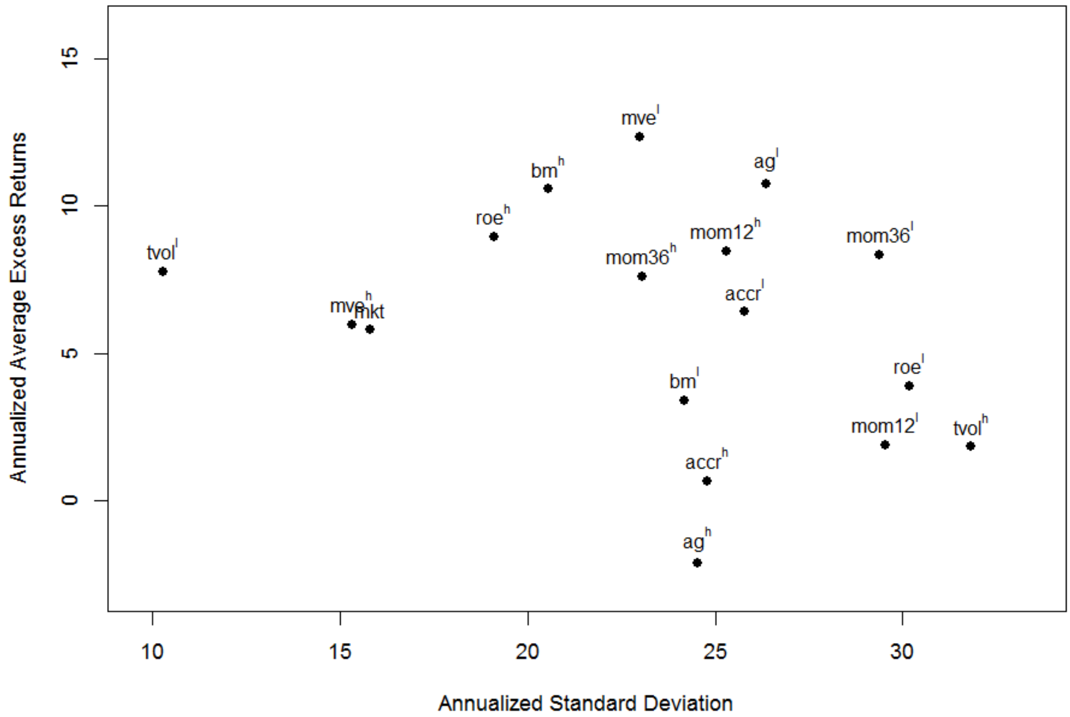
Romano, J. P. and Wolf, M. (2007). "Control of generalized error rates in multiple testing." *Annals of Statistics*, **35**, 1378-1408.

Stattman, D. (1980). "Book values and stock returns." *The Chicago MBA: A Journal of Selected Papers*, **4**, 25-45.

Sloan, R. G. (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings." *The Accounting Review*, **71**, 289-315.

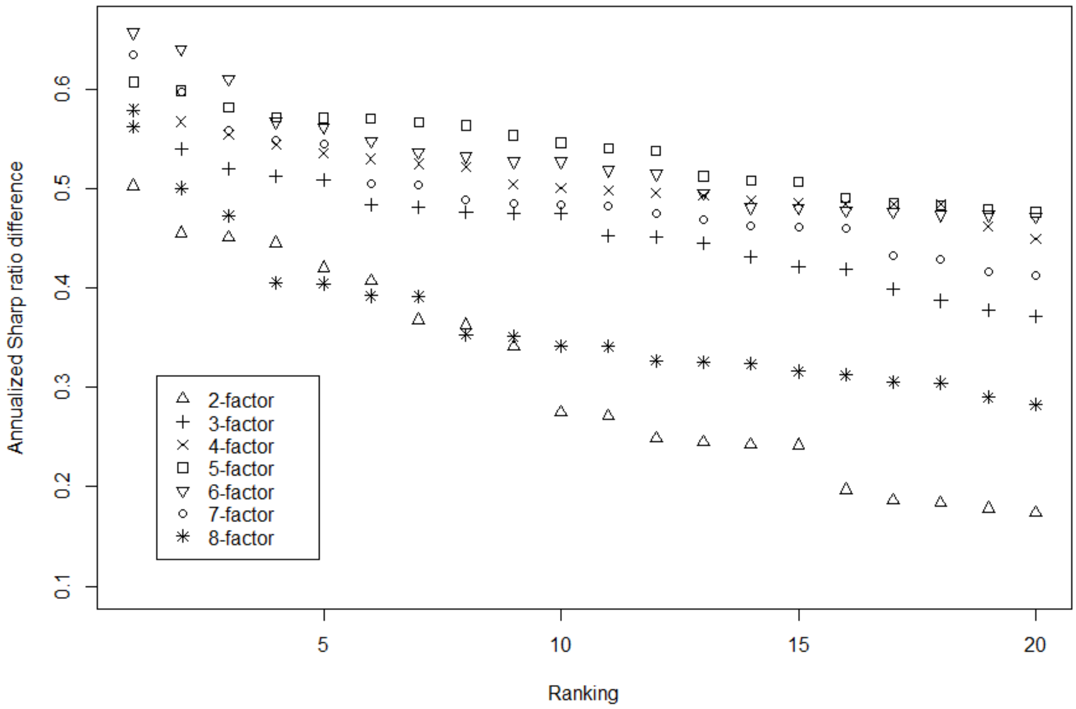
Wigglesworth, R. "Fund managers ready for smart beta wars." *Financial Times*, February 8, 2016. http://on.ft.com/1SDFAub (accessed March 21, 2016).

Figure 1. Mean excess returns and standard deviation.



Note: The mean and standard deviation are computed with score-weighted portfolio returns from May 1968 to December 2015. The portfolio formation date is at the end of April every year. The superscripts *h* and *l* denote the high and low decile portfolios, respectively. The portfolio *mkt* is the value-weighted market portfolio.

Figure 2. Sharpe ratio differences comparison.



Note:This figure provides a comparison of the Sharpe ratio differences among the top 20 portfolios across different number of factors used in each multi-factor strategy.

Table 1. Variable definition.

Style	Factor	Reference	Definition
Value	*bm*	Stattman (1980)	book value of equity / market value of equity
Size	*mve*	Banz (1981)	market value of equity
Momentum	*mom12*	Jegadeesh (1990)	cumulative stock returns over the past twelve months, excluding the most recent month
Reversal	*mom36*	DeBondt and Thaler (1985)	cumulative stock returns over the past three years
Risk	*tvol*	Ang et al. (2006)	historical volatility of stock returns over the past 52 weeks
Profitability	*roe*	Fama and French (2006)	earnings / book value of equity
Growth	*ag*	Cooper et al. (2008)	annual growth of total assets
Earnings Quality	*accr*	Sloan (1996)	(EBIT – cash flow from operations) / beginning-of-year total assets

Note: The portfolio formation time is end of April each year. Annual accounting information is assumed to be available with a four-month lag and only firms with December fiscal year end are considered. Book value of equity excludes preferred stocks.

Table 2. Sharpe ratio difference of multifactor strategies.

Number of Factors	Number of Portfolios	Number of Outperforming Strategies	Critical Value for Sharpe Ratio Difference
1	16	1	0.240
2	112	11	0.255
3	448	61	0.242
4	1120	126	0.246
5	1792	196	0.246
6	1792	197	0.246
7	1024	104	0.248
8	256	25	0.250

Note: The critical value of. Sharpe ratio difference is defined in Equation (2). The annualized Sharpe ratio of market portfolio is 0.366. The number of portfolios in each universe is $C_j^8 \times 2^j$, where *j* is the number of factors used to construct the multi-factor portfolios.

Table 3. Strategies with the highest Sharpe ratio difference.

Number of Factors	Top 5 Strategies
1	$tvol^l$ (0.391), mve^l (0.170), bm^h (0.147), roe^h (0.102), ag^l (0.041).
2	$mve^l/tvol^l$ (0.502), $mom12^h/tvol^l$ (0.455), $bm^h/tvol^l$ (0.450), $tvol^l/ag^l$ (0.445), $mom36^h/tvol^l$ (0.420).
3	$mve^l/mom12^h/tvol^l$ (0.578), $mve^l/mom36^h/tvol^l$ (0.540), $mve^l/tvol^l/roe^h$ (0.520), $bm^h/mom12^h/tvol^l$ (0.512), $bm^h/mom36^h/tvol^l$ (0.508).
4	$bm^h/mve^l/mom12^h/tvol^l$ (0.579), $bm^h/mve^l/mom36^h/tvol^l$ (0.567), $bm^h/mom36^h/tvol^l/ag^l$ (0.554), $mve^l/mom12^h/tvol^l/roe^h$ (0.544), $bm^h/mom12^h/tvol^l/ag^l$ (0.536).
5	$bm^h/mve^l/mom12^h/tvol^l/roe^h$ (0.607), $bm^h/mve^l/mom36^h/tvol^l/ag^l$ (0.598), $mve^l/mom36^h/tvol^l/ag^l/roe^h$ (0.581), $mve^l/mom12^h/tvol^l/ag^l/roe^h$ (0.571), $bm^h/mve^l/mom12^h/mom36^h/tvol^l$ (0.571).
6	$bm^h/mve^l/mom12^h/tvol^l/ag^l/roe^h$ (0.657), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h$ (0.640), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l$ (0.610), $bm^h/mve^l/mom36^h/tvol^l/ag^l/accr^l$ (0.567), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/roe^h$ (0.561).
7	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h$ (0.634), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.597), $bm^h/mve^l/mom12^h/tvol^l/ag^l/roe^h/accr^l$ (0.559), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/roe^h/accr^l$ (0.548), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/accr^l$ (0.545).
8	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.562), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h/accr^h$ (0.500), $bm^h/mve^l/mom12^l/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.472), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^l/accr^h$ (0.405), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^l/accr^l$ (0.404).

Note: The number in parentheses is the portfolio's Sharpe ratio difference.

Table 4. In-sample and out-of-sample comparison.

Panel A.

Number of Factors	Ranked with In-Sample Data	Ranked with Out-of-Sample Data
1	$tvol^l$ (0.367, 0.650)	$tvol^l$ (0.367, 0.650)
	bm^h (0.255, -0.753)	mve^h (0.037, -0.115)
	mve^l (0.228, -0.307)	$mom36^h$ (-0.013, -0.194)
2	$mve^l/tvol^l$ (0.491, 0.865)	$mve^l/tvol^l$ (0.491, 0.865)
	$bm^h/tvol^l$ (0.466, 0.367)	$mom36^h/tvol^l$ (0.403, 0.653)
	$tvol^l/ag^l$ (0.452, 0.359)	$mom12^h/tvol^l$ (0.444, 0.522)
3	$mve^l/mom12^h/tvol^l$ (0.565, 0.795)	$mve^l/mom36^h/tvol^l$ (0.530, 0.904)
	$mve^l/tvol^l/roe^h$ (0.532, 0.595)	$mve^l/mom12^h/tvol^l$ (0.565, 0.795)
	$mve^l/mom36^h/tvol^l$ (0.530, 0.904)	$mve^l/tvol^l/accr^h$ (0.167, 0.614)
4	$bm^h/mve^l/mom12^h/tvol^l$ (0.585, 0.616)	$bm^h/mve^l/mom36^h/tvol^l$ (0.564, 0.857)
	$bm^h/mom36^h/tvol^l/ag^l$ (0.565, 0.487)	$mve^l/mom36^h/tvol^l/ag^l$ (0.529, 0.682)
	$bm^h/mve^l/mom36^h/tvol^l$ (0.564, 0.857)	$mve^l/mom36^h/tvol^l/roe^l$ (0.531, 0.643)
5	$bm^h/mve^l/mom12^h/tvol^l/roe^h$ (0.620, 0.593)	$bm^h/mve^l/mom36^h/tvol^l/ag^l$ (0.610, 0.682)
	$bm^h/mve^l/mom36^h/tvol^l/ag^l$ (0.610, 0.682)	$mve^l/mom36^h/tvol^l/ag^l/accr^h$ (0.323, 0.649)
	$mve^l/mom36^h/tvol^l/ag^l/roe^h$ (0.591, 0.619)	$bm^h/mve^l/mom12^h/mom36^h/tvol^l$ (0.579, 0.632)

Panel B.

Number of Factors	Number of Portfolios	Number of Outperforming Strategies	Critical Value for Sharpe Ratio Difference
1	16	1	0.264
2	112	14	0.269
3	448	71	0.262
4	1120	136	0.267
5	1792	237	0.265

Notes: Panel A presents the top 3 outperforming multifactor strategies based on in-sample and out-of-sample period ranking. We use May 1968 to April 2010 as the in-sample period and May 2010 to December 2015 as the out-of-sample period. The numbers in the parentheses are Sharpe ratio differences for the in-sample period (first) and out-of-sample period (second). The Sharpe ratios of market portfolio are 0.304 and 0.912 for the in-sample and out-of-sample period, respectively. Panel B reports the critical values estimated using only the in-sample period data.

Table 5. Backtesting with small-cap value portfolio as the benchmark.

Number of Factors	Number of Outperforming Strategies	Critical Value for Sharpe Ratio Difference	Top 3 Significantly Outperforming Strategies
3	10	0.286	$mve^l/mom12^h/tvol^l$ (0.410), $mve^l/mom36^h/tvol^l$ (0.372), $mve^l/tvol^l/roe^h$ (0.352).
4	18	0.300	$bm^h/mve^l/mom12^h/tvol^l$ (0. 412), $bm^h/mve^l/mom36^h/tvol^l$ (0.399), $bm^h/mom36^h/tvol^l/ag^l$ (0.386).
5	24	0.301	$bm^h/mve^l/mom12^h/tvol^l/roe^h$ (0.439), $bm^h/mve^l/mom36^h/tvol^l/ag^l$ (0.431), $mve^l/mom36^h/tvol^l/ag^l/roe^h$(0.414).
6	21	0.296	$bm^h/mve^l/mom12^h/tvol^l/ag^l/roe^h$ (0.489), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h$ (0.472), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l$ (0.442).
7	14	0.294	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h$ (0.467), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.430), $bm^h/mve^l/mom12^h/tvol^l/ag^l/roe^h/accr^l$ (0.391).
8	3	0.291	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.394), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h/accr^h$ (0.333), $bm^h/mve^l/mom12^l/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.305).

Notes: This table summarizes the backtesting results when small-cap value portfolio is used as the benchmark. For the last column, the number in parentheses is the Sharpe ratio difference between the multi-factor portfolio and small-cap value portfolio (bm^h/mve^l). The Sharpe ratio of small-cap value portfolio in our sample period is 0.534.

Table 6. Results using alternative proxy variable.

Panel A.

Style	Factor	Definition
Value	*ey*	EBIT / market value of equity
Size	*mve*	market value of equity
Momentum	*wh52*	current stock price / 52-week high of stock price
Reversal	*mom60*	cumulative stock returns over the past sixty months
Risk	*beta*	CAPM beta from regression using the past 52 weekly returns
Profitability	*gpm*	gross profit margin
Growth	*inv*	capital expenditure / gross property, plant, and equipment
Earnings Quality	*accr*	(EBIT – cash flow from operations) / last year total assets

Panel B.

Number of factors	Benchmark: Market portfolio		Benchmark: Small-cap value portfolio	
	Critical value	Number of outperforming strategies	Critical value	Number of outperforming strategies
1	0.236	2	0.163	0
2	0.241	21	0.216	0
3	0.240	95	0.257	2
4	0.235	245	0.273	4
5	0.234	392	0.275	3
6	0.232	410	0.271	1
7	0.230	246	0.260	0
8	0.233	63	0.232	0

Panel C.

	List of the best 5 strategies
Single Factor	$wh52^h$ (0.404), ey^h (0.352), mve^l (0.170), $beta^l$ (0.150), ag^l (0.147).
Multi-Factor	$ey^h/mve^l/wh52^h/beta^l$ (0.758), $ey^h/mve^l/wh52^h/beta^l/gpm^h$ (0.736), $ey^h/mve^l/wh52^h$ (0.717), $ey^h/wh52^h/beta^l$ (0.712), $ey^h/mve^l/wh52^h/gpm^h$ (0.702).

Notes: This table presents the results with alternative factor definitions. Panel A provides the definition of each variable. Panel B shows the critical values for Sharpe ratio differences and the number of outperforming strategies. The Sharpe ratios of market portfolio and small-cap value portfolio (ey^h/mve^l) are 0.366 and 0.407, respectively. In Panel C, we report the five portfolios with the highest Sharpe ratio for single factor and multi-factor portfolios; the number in parentheses is the Sharpe ratio difference between the corresponding portfolio and market portfolio.

Table 7. Value-weighted portfolio.

Number of Factors	Number of Outperforming Strategies	Critical Value for Sharpe Ratio Difference	The Top 3 Portfolios with the Highest Sharpe Ratios
1	0	0.218	$tvol^{l}$ (0.149), bm^{h} (0.14), ag^{l} (0.132).
2	2	0.276	$mve^{l}/tvol^{l}$ (0. 407), $tvol^{l}/ag^{l}$ (0.302), $bm^{h}/tvol^{l}$ (0.206).
3	11	0.281	$mve^{l}/mom12^{h}/tvol^{l}$ (0. 504), $mve^{l}/mom36^{h}/tvol^{l}$ (0.463), $mve^{l}/tvol^{l}/roe^{h}$ (0.453).
4	16	0.297	$bm^{h}/mve^{l}/mom36^{h}/tvol^{l}$ (0. 483), $bm^{h}/mve^{l}/mom12^{h}/tvol^{l}$ (0.477), $mve^{l}/mom12^{h}/tvol^{l}/ag^{l}$ (0.459).
5	26	0.292	$bm^{h}/mve^{l}/mom12^{h}/tvol^{l}/ag^{l}$ (0.479), $bm^{h}/mve^{l}/tvol^{l}/ag^{l}/roe^{h}$ (0.424), $bm^{h}/mve^{l}/mom36^{h}/tvol^{l}/ag^{l}$ (0.421).
6	25	0.293	$bm^{h}/mve^{l}/mom36^{h}/tvol^{l}/ag^{l}/roe^{h}$ (0.463), $bm^{h}/mve^{l}/mom12^{h}/ mom36^{h}/tvol^{l}/ag^{l}$ (0.451), $bm^{h}/mve^{l}/mom36^{h}/tvol^{l}/ag^{l}/accr^{l}$ (0.450).
7	16	0.296	$bm^{h}/mve^{l}/mom36^{h}/tvol^{l}/ag^{l}/roe^{h}/accr^{l}$ (0.433), $bm^{h}/mve^{l}/mom12^{h}/mom36^{h}/tvol^{l}/ag^{l}/roe^{h}$ (0.428), $bm^{h}/mve^{l}/mom12^{h}/tvol^{l}/ag^{l}/roe^{h}/accr^{l}$ (0.422).
8	4	0.292	$bm^{h}/mve^{l}/mom12^{h}/mom36^{l}/tvol^{l}/ag^{l}/roe^{h}/accr^{l}$ (0.373), $bm^{h}/mve^{l}/mom12^{h}/mom36^{h}/tvol^{l}/ag^{l}/roe^{l}/accr^{h}$ (0.352), $bm^{h}/mve^{l}/mom12^{h}/mom36^{h}/tvol^{l}/ag^{l}/roe^{h}/accr^{h}$ (0.343).

Notes: This table shows the results when market capitalization is used to determine the portfolio weight. The number in parentheses is the Sharpe ratio difference between the portfolio and market portfolio.

Tables not for publication.

Table A1. Each factors is trimmed at 5%.

Number of Factors	Number of Portfolios	Number of Outperforming Strategies	Critical Value of Sharpe Ratio Difference
1	16	1	0.233
2	112	13	0.248
3	448	76	0.231
4	1120	174	0.240
5	1792	302	0.235
6	1792	315	0.235
7	1024	170	0.240
8	256	46	0.240

Note: Each factors is trimmed using 5% level before constructing the combined score. The critical value of. Sharpe ratio difference is defined in Equation (2) annualized Sharpe ratio Market portfolio's Sharpe ratio is 0.366. The number of portfolios in each universe is $C_j^8 \times 2^j$, where j is the number of factors used to construct the multi-factor portfolios.

Table A2. Top 5 outperforming strategies with 5% trimming.

Number of Factors	Top 5 Strategies
1	$tvol^l$ (0.381), ag^l (0.171), bm^h (0.142), roe^h (0.107), $mom36^h$ (0.081).
2	$mom12^h/tvol^l$ (0.433), $mve^l/tvol^l$ (0.401), $bm^h/tvol^l$ (0.397), $mom36^h$ $/tvol^l$ (0.397), $tvol^l/ag^l$ (0.393) .
3	$mve^l/mom12^h$ $/tvol^l$ (0.523), $bm^h/mom36^h/tvol^l$ (0.517), $bm^h/mom36^h/ag^l$ (0.503), $mve^l/mom36^h$ $/tvol^l$ (0.496), $bm^l/mom12^h/tvol^l$ (0.480).
4	$bm^h/mom36^h/tvol^l/ag^l$ (0.531), $mve^l/mom36^h/tvol^l/accr^l$ (0.523), $mve^l/mom36^h/tvol^l/ag^l$ (0.515), $bm^h/mve^l/mom12^h/tvol^l$ (0.514), $bm^h/mom12^h/tvol^l/ag^l$ (0.512).
5	$mve^l/mom12^h/tvol^l/ag^l/roe^h$ (0.571), $bm^h/mom36^h/tvol^l/ag^l/accr^l$ (0.565), $bm^h/mve^l/mom36^h/tvol^l/accr^l$ (0.551), $bm^h/mve^l/mom36^h/tvol^l/ag^l$ (0.536), $bm^h/mve^l/mom12^h/mom36^h/tvol^l$ (0.526).
6	$bm^h/mve^l/mom12^h/mom36^h$ $/tvol^l/ag^l$ (0.567), $bm^h/mve^l/mom12^h/tvol^l/ag^l/roe^h$ (0.563), bm^h/mve^l $/mom36^h/tvol^l/ag^l/accr^l$ (0.563), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/accr^l$ (0.540), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h$ (0.535).
7	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h$ (0.558), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/accr^l$ (0.553), $bm^h/mve^l/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.535), $bm^h/mve^l/$ $mom12^h/tvol^l/ag^l/roe^h/accr^l$ (0.511), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/roe^h/accr^l$ (0.493).
8	$bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.540), $bm^h/mve^l/mom12^h$ $/mom36^h/tvol^l/ag^l/roe^l/accr^l$ (0.488), $bm^h/mve^l/mom12^l$ $/mom36^h/tvol^l/ag^l/roe^h/accr^l$ (0.464), $bm^h/mve^l/mom12^h/mom36^h/tvol^h/ag^l/roe^h/accr^l$ (0.433), $bm^h/mve^l/mom12^h/mom36^h/tvol^l/ag^l/roe^l/accr^h$ (0.424).

Note: The number in parentheses is the portfolio's Sharpe ratio difference.

Table A3. Portfolio with fixed number of stocks.

	Number of factors	1	2	3	4	5	6	7	8
S = 50	Critical value	0.283	0.331	0.298	0.294	0.298	0.289	0.291	0.281
	No. of outperforming strategies	1	7	37	84	114	142	74	24
S = 100	Critical value	0.270	0.272	0.271	0.274	0.267	0.265	0.259	0.252
	No. of outperforming strategies	1	11	45	88	161	183	107	30
S = 200	Critical value	0.243	0.254	0.249	0.251	0.245	0.243	0.238	0.232
	No. of outperforming strategies	1	12	51	101	197	212	127	35
S = 300	Critical value	0.238	0.244	0.237	0.239	0.236	0.232	0.222	0.216
	No. of outperforming strategies	1	21	51	108	216	230	154	42

Note: The portfolios are score-weighted of fixed number selected *S* stocks and rebalanced every year.