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

**Kendro Vincent**
Department of International Business
National Taiwan University

**Yu-Chin Hsu**
Institute of Economics
Academia Sinica

**Hsiou-Wei Lin**
Department of International Business
National Taiwan University
Department of International Business
Tunghai 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 from serious data mining bias and lead to spurious significant findings. Accordingly, we employ a multiple hypothesis testing method to examine the multi-factor portfolios performance. Our empirical results show that even after we adjust for the multiple comparisons bias, the stock picking strategies with certain combined firm characteristics could generate significantly better liquidity risk-adjusted returns. In addition, the outperforming multi-factor strategies are robust to alternative definitions of factors. However, we observe that the number of significantly profitable multi-factor portfolios decreases substantially in the era of increased liquidity and trading activity in the U.S. stock market.

*Keywords*: Data mining bias, multi-factor investment strategy, multiple hypothesis testing, passive index investing, smart beta.
*JEL classifications*: G11, G17.

## Introduction

The ETFs under the label of *smart beta* or *factor investing* 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 a single-factor, 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*, see e.g. Authers [2015], Noblett [2015], and Wigglesworth [2016].

Skeptics have been suggesting that the *smart beta* strategies are merely smart marketing without any value being added for investors. For example, Malkiel [2014] criticizes the *smart beta* ETFs for their not consistently outperforming the respective benchmark. Arnott et al. [2013] also suggest that multi-factor strategies resemble small-cap value portfolio. Namely, they simply provide product varieties without gaining risk-adjusted returns. On the other hand, several papers have pointed out that there may exist a bias in the traditional testing methodology induced by the multiplicity in backtesting; see for example McQueen and Thorley [1999], Bailey and de Prado [2014], Harvey and Liu [2014], and Novy-Marx [2016].

In this article, we attempt to answer the skeptics by testing the multi-factor investment strategies on U.S. stock data over the sample period between 1973 and 2016. The investigation throughout this study is conducted under multiple testing framework to resolve the data mining bias issue. We adopt a data-driven method by Hsu et al. [2014] to control the number of false rejections in testing the performance of the top among a large number of portfolios. Moreover, to take the correlations among portfolios into account, the multiple testing algorithm uses re-sampling distribution in obtaining the critical value that separates the superior and inferior multi-factor portfolios. This is particularly useful since the multi-factor portfolios are expected to be correlated. For example, value-momentum and profitability-momentum strategies may be interrelated because of the common exposure to the momentum factor.

In addition to mitigating the issue of multiple comparisons, we base our portfolio performance evaluation on the empirical asset pricing model proposed by Pastor and Stambaugh [2003], in which the abnormal returns calculation is adjusted for illiquidity risk. The robustness of multi-factor portfolio results is then explored by testing with different subsamples and factors definition.

## Overview of Multiple Testing

In identifying the portfolio with superior performance, an investment manager rarely evaluates only one portfolio. In fact, one of the common tasks in investment management is examining the performance of a large set of portfolios repeatedly. More formally, the investment manager faces the problem of testing multiple inequalities,

$$H^i: \theta_i \leq 0, \qquad i = 1, \dots, \mathrm{M}, \tag{1}$$

where M is the number of portfolios considered and $\theta_i$ measures the performance of portfolio $i$. If the individual hypothesis testing with 5% significance level is used, then one out of twenty portfolios could show promising outperformance with high probability even though all of the true $\theta_i$ are less than zero.[1]

To avoid the multiple hypothesis testing (multiple comparisons) bias, the statistical inference should be conducted under one family of hypotheses and controls for a different notion of error rate rather than the conventional Type I error. One of the commonly used error rates in the multiple testing literature is $k$-family-wise error rate ($k$-FWER), the probability of rejecting at least $k$ true hypotheses $H^i$. One of the procedures to control $k$-FWER is the Step-SPA($k$) algorithm by Hsu et al. [2014] that incorporates dependence information via bootstrap.

To see how the Step-SPA($k$) algorithm works, suppose we have a threshold $\boldsymbol{c}$ for the test statistics of the performance measures that helps differentiate the superior portfolios from the inferior ones. Note that we commit $k$ false rejections only if the $k$-th largest value of the test statistics of the inferior portfolios exceeds our specified threshold $\boldsymbol{c}$. The

[1] Suppose all of the twenty portfolios have zero $\theta$ and are independent with one another. The probability of finding one spurious outperforming portfolio is equal to one minus the probability that all $\theta$'s are not rejected simultaneously, $1 - (1 - 0.05)^{20} = 0.64$.

objective of a multiple testing algorithm is to find the threshold $\boldsymbol{c}$ such that $k$-FWER is bounded by a small probability $\delta$. In a formal notation,

$$k\text{-FWER} = \text{P}\{\text{ The } k\text{-th largest value of } \sqrt{T}(\hat{\theta}_i - \theta_i) > \boldsymbol{c}, i \in \mathbf{H_0} \} \leq \delta, \tag{2}$$

where $T$ denotes the sample size to compute the consistent estimator $\hat{\theta}_i$ and $\mathbf{H_0}$ denotes the set of portfolios with true $\theta_i \leq 0$.

The idea behind Step-SPA($k$) algorithm is fairly intuitive. First, it uses the resampling approach to approximate the distribution of the necessary order statistics, so that the critical value for the $k$-th largest test statistics could be obtained by the bootstrap samples accordingly. Second, since the set $\mathbf{H_0}$ is unknown, Step-SPA($k$) starts with the most conservative guess by using $\mathbf{H_0} = \{1, \dots, \text{M}\}$ to derive an initial threshold. If there is a non-empty subset of the portfolios $\mathbf{R} \subseteq \{1, \dots, \text{M}\}$ being rejected, a new critical value is obtained with the refined set $\mathbf{H_0'}$, which contains $\mathbf{H_0} \backslash \mathbf{R}$ and $k$-1 of the least significant portfolios from $\mathbf{R}$. The inclusion of $k$-1 portfolios from $\mathbf{R}$ is necessary to constrain the procedure from inflating $k$-FWER.[2] The stepwise procedure is repeated until there is no further rejection, and the last critical value is then obtained as the threshold $\boldsymbol{c}$ in (2).

The appropriate choice of $k$ in controlling $k$-FWER may vary from case to case. For instance, a small number of $k$ may appear too restrictive if a portfolio manager chooses from a larger pool of investment opportunities to diversify. Therefore, it is desirable to control False Discovery Proportion (FDP), which is defined as the ratio of the number of false rejections to the total number of rejections.[3] To prevent the performance evaluation from finding too many spurious outperforming strategies, we control the probability of FDP exceeding a level $\xi$ with a small probability $\delta$,

$$\text{P}\{\text{FDP} > \xi\} \leq \delta. \tag{3}$$

[2] To see why it is necessary to include $k$-1 portfolios from previous rejections, recall that the initial threshold is derived so that P{number of false rejections $\geq k$} is less than $\delta$, which also implies that P{number of false rejections $\leq k$-1} is at least 1-$\delta$. In other words, we could have at most $k$-1 false rejections in the previous rejections with high probability. Therefore, the refined set $\mathbf{H_0'}$ must conform to this to ensure the overall $k$-FWER is still controlled below α. For the implementation details, please see the "while" loop (line 21) of the pseudo-code for Step-SPA($k$) algorithm in the Appendix.

[3] If there is no rejection, then FDP is defined as 0.

In our empirical studies, we aim to achieve the FDP exceedance control by adopting the FDP-SPA recursive algorithm outlined in Hsu et al. [2014]. The FDP-SPA method proceeds as follows. We begin from Step-SPA(1) at significance level $\delta$ and reject the portfolios with performance metrics being greater than the estimated critical value. If there is no rejection, i.e. none of the portfolios have 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 exceedance below $\delta$. The exact algorithms of the Step-SPA($k$) and FDP-SPA are given in the Appendix.

The FDP exceedance control is related to the False Discovery Rate (FDR), defined as the expected value of FDP, controlling procedure suggested by Harvey and Liu [2014] and Harvey et al. [2016]. If the test statistics are correlated, then the distribution of realized FDP could become less concentrated and skewed, see Figure 1 of Delattre and Roquain [2015] for the simulation results. As a consequence, the true value of FDP may not be bounded below a small value. The results in this study can be viewed as more conservative in the sense that FDP exceedance control also guarantees FDR control. The inequality (3) implies that FDR is controlled below $\xi + (1 - \xi)\delta$. In the empirical analyses, we set both $\xi$ and $\delta$ to be 0.05, hence the implied FDR rate is also controlled at 9.75%. The level of FDR control is usually suggested to be 10%.

The multiple testing of portfolio performance with Step-SPA($k$) and FDP-SPA can be based on the studentized statistic, $\sqrt{T}(\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$ (the $t$-statistic) or the non-studentized one, $\sqrt{T}(\hat{\theta}_i - \theta_i)$. In this study, if the alpha is used as the portfolio performance measure, we use the studentized statistic so that the comparison is on the same scale. For the analysis with Sharpe ratio, we simply use the non-studentized statistic.

## Measuring Portfolio Performance

Following Fama and French [1992, 1993] and Carhart [1997], we use the linear risk factor model to compute the risk-adjusted returns or alpha. Specifically, the alpha of portfolio $i$ is the estimated intercept coefficient of the time-series regression:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_{1,i}\left(r_{m,t} - r_{f,t}\right) + \beta_{2,i}SMB_t + \beta_{3,i}HML_t + \beta_{4,i}UMD_t + \varepsilon_{i,t}\,, \qquad (4)$$

where $r_{i,t}$, $r_{m,t}$, and $r_{f,t}$ are the portfolio returns, value-weighted market portfolio returns, and risk-free rate, respectively. The additional risk factors *SMB*, *HML*, and *UMD* play the role of removing the systematic component of portfolio risk other than market-wide risk. The *t*-ratio is calculated by dividing $\hat{\alpha}_i$ with Newey-West standard error, which is robust to autocorrelation in $\varepsilon_{i,t}$. Moreover, we augment the set of risk factors in (4) with traded liquidity factor *LIQ* by Pastor and Stambaugh [2003] to capture the time-varying illiquidity risk.[4] We also measure the performance of portfolio *i* by estimating its annualized Sharpe ratio difference between the portfolio and the market portfolio, $SR_i - SR_m$. The annualized Sharpe ratio is defined as $\sqrt{12}\mathbf{E}(r_{i,t} - r_{f,t})/\boldsymbol{\sigma}(r_{i,t} - r_{f,t})$.

In addition to evaluating the portfolio risk-adjusted returns, we analyze the portfolio turnover and transaction cost. While it is straightforward to compute the portfolio turnover, we can only find imperfect transaction cost measure which could incorporate price impact as well. We follow Acharya and Pedersen [2005] to construct a per-unit trade cost for each stock *j* as

$$tc_{j,t} = \min\left(0.25 + 0.3\, illiq_{j,t}\, PM_{t-1}\,, 30\right), \tag{5}$$

where $PM_{t-1}$ is the ratio of the capitalization of the market portfolio at the end of month *t*-1 and that of the market portfolio at the end of July 1962, and $illiq_{j,t}$ is the illiquidity measure of stock *j* at the end of month *t* by Amihud [2002], defined as

$$illiq_{j,t} = \frac{1}{Days_{j,t}} \sum_{d=1}^{Days_{j,t}} \frac{|r_{j,d}|}{DV_{j,d}}, \tag{6}$$

where $r_{j,d}$ and $DV_{j,d}$ are respectively the daily return and dollar volume in millions on day *d* in month *t*, and $Days_{j,t}$ is the number of trading days in month *t*. Goyenko et al. [2009] perform a series of horse race evaluations and conclude that $illiq_{j,t}$ appears to be the winner among the considered proxies for measuring illiquidity and price impact.

The adjustment factor $PM_{t-1}$ is used to eliminate the inflation effect in the denominator of $illiq_{j,t}$. Furthermore, Acharya and Pedersen [2005] choose the coefficients in (5) to

[4] There are two versions of traded liquidity factors in Pastor and Stambaugh [2003]: value-weighted and equal-weighted. We adopt the value-weighted one, which is in line with the conventional method to construct risk factors such as Fama and French [1992, 1993] and Carhart [1997].

calibrate the mean effective spread so that $tc_{j,t}$ can be interpreted as per-unit trade cost. In this setting, the cost of trading a stock is at least 0.25% and would be greater depending on the price impact. We calculate the transaction cost of portfolio *i* as

$$TC_{i,y} = \sum_{j=1}^{N} |w_{j,y}^{new} - w_{j,y}^{old}| \, tc_{j,y} \, , \tag{7}$$

where $w_{j,y}^{new}$ and $w_{j,y}^{old}$ are, respectively, the new and old portfolio weights of stock *j* at the time of rebalancing.[5]

## Data and Portfolio Construction

We obtain the accounting and monthly stock returns data of US firms over 1973-2016 from Compustat and CRSP databases, respectively. The data for risk factors and risk-free rate are provided by WRDS. We follow Beaver et al. [2007] to adjust for delisting bias in CRSP stock returns data. A firm's stock is eligible for portfolio construction process if at the time of portfolio formation: (1) it has positive book value of equity; (2) it belongs to non-financial industry; (3) it has December fiscal year end; (4) its stock price ranges between $5 and $1,000; (5) and its market capitalization is greater than the 501-st largest in the universe of stocks after filtering rules (1)-(4). The conditions (4) and (5) follow Acharya and Pedersen [2005] and Amenc et al. [2016], respectively. Both selection criteria help ensure that we do not include too many illiquid stocks in the portfolios.

Over the past few decades, the empirical finance literature has documented quite a few factors affecting stock returns, Cochrane [2011] dubs this phenomena zoo of factors. We categorize the factors into 8 styles. Exhibit 1 provides the factors definition. While each style is already examined and supported in the literature, the best factor for each style is not known ex-ante. We pick one representative variable for each style as the factor to avoid increasing number of combinations and deterioration in the power of statistical test.[6] For

[5] As explained in the next section, we construct a setting that reconstitute and rebalance the portfolio annually at end of June. Here $tc_{j,y}$ is the transaction cost at the end of June of year *y*. The old portfolio weight, $w_{j,y}^{old}$, is determined by the portfolio weight from last year $w_{j,y-1}^{new}$ and the cumulative gross return during the one-year holding period. Also note that the transaction cost is only incurred once a year. Likewise, the calculation of portfolio turnover $\sum_{j=1}^{N} |w_{j,y}^{new} - w_{j,y}^{old}|$ follows the same rule.

[6] Intuitively, the critical value would be greater as the number of portfolios in $\mathbf{H_0}$ increases. Here we avoid inflating the number of portfolios with same style on purpose, thus decreasing the power of finding superior portfolios.

robustness check, we construct portfolios by using different measurements of the style factors.

An anecdotal survey of multi-factor index, summarized in Exhibit 2, shows that the scope of our factors coverage should be broad enough. The common theme of the multi-factor index is the inclusion of size, value, and momentum. The additional factors to combine vary across different indexes. We also include other influential stock return determinants documented in the academic literature, such as accrual quality (Sloan [1996] and long-term reversal (DeBondt and Thaler [1985]), to provide a more comprehensive analysis despite that these factors are not covered in the major multi-factor index.

Exhibit 1. Factors Definition.

| Style | Factor | Definition |
|---|---|---|
| Size | *mve* | Market value of equity. |
| Value | *bm* | Book value of equity / market value of equity. |
| Profitability | *roe* | Earnings / book value of equity. |
| Investment | *ag* | Annual growth of total assets. |
| Momentum | *mom12* | Cumulative stock returns over the past twelve months, excluding the most recent month. |
| Long-term Reversal | *rev36* | Cumulative stock returns over the past three years, excluding the most recent year. |
| Risk | *tvol* | Historical volatility of weekly stock returns over the past 52 weeks |
| Accrual Quality | *accr* | (EBIT – cash flow from operations) / beginning-of-year total assets. |

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

The construction of the multi-factor portfolios is as follows. At the end of June every year, a percentile score variable *scr*(*x*) is created for each factor *x*. Procedurally, we rank *bm*, *roe*, and *mom12* from high to low and assign *scr*(*bm*), *scr*(*roe*), and *scr*(*mom12*) the values 1, $\frac{499}{500}$, …, $\frac{1}{500}$ based on their rankings.[7] For *mve*, *ag*, *rev36*, *tvol*, and *accr*, the factors are ranked in an ascending order and then transformed into 1, $\frac{499}{500}$, …, $\frac{1}{500}$. The multi-factor stock selection picks the top 20% of the stocks based on the combined score,

[7] If there are ties, we assign the mean score. For example, suppose there are only 5 stocks and *roe* = (0.3, 0.2, 0.2, 0.12, 0.06), then *scr*(*roe*) = (1, $\frac{3.5}{5}$, $\frac{3.5}{5}$, $\frac{2}{5}$, $\frac{1}{5}$).

which results in holding approximately 100 stocks in the portfolio. For instance, the long-only value-momentum strategy involves buying the stocks within top 20% of *scr*(*bm*) + *scr*(*mom12*).[8] We focus on the long-only strategy in the main empirical analysis because this is what the multi-factor index ETFs mostly adopt.

Exhibit 2. Examples of Multi-Factor Index.

| Index | Styles |
|---|---|
| MSCI Diversified Multi-Factor | Size, value, momentum, and quality*. |
| John Hancock Dimensional Large-Cap | Size, value, momentum, and profitability. |
| FTSE Diversified Factor | Size, value, momentum, and volatility |
| Scientific Beta Multi-Factor Equal Risk Contribution | Size, value, momentum, volatility, profitability, and asset growth. |

*The quality measure is a combination of ROE, debt-to-equity ratio, and earnings volatility.

We consider three types of portfolio weighting schemes: value, equal, and score weighting. The value-weighted portfolio weights each selected stocks with its market-capitalization at the end of June, and the equal-weighted portfolio invests in each selected stocks with equal amount of dollar. We follow a similar approach by Asness et al. [2013] and Novy-Marx [2016] to construct score-weighted portfolio by weighting each selected stocks with

$$w_{j,t} = v_t |scr_{j,t} - Q_t^q(scr)|, \tag{8}$$

where $scr_{j,t}$ is the combined score, $Q_t^q(scr)$ is the cross-sectional upper $q$% quantile of $scr_{j,t}$, and $v_t$ is the normalized constant such that the portfolio weights sum to one. The portfolio is rebalanced at the end of June each year.

## Empirical Analyses

Before diving into the analysis of multi-factor portfolios, let us review the results of single-factor portfolios for a reference on how combining more than one factors could lead to outperformance. Moreover, we demonstrate how the statistical inference based on multiple testing leads to different conclusions from the individual testing.

[8] This approach follows Stambaugh and Yu [2016]. A similar method is used by FTSE Diversified Multi-Factor index, MSCI Diversified Multi-Factor index, Dimensional Large-Cap index, Greenblatt [2010], and Novy-Marx [2016].

Exhibit 3. Long-Only Portfolios Formed by the Top 20% of Single-Factor.

| | Excess Return | Sharpe Ratio Difference | 4-Factor Alpha | 4-Factor+*LIQ* Alpha | Turnover | Transaction Cost |
|---|---|---|---|---|---|---|
| *A.* Value weighting | | | | | | |
| *mve* | 11.28 | 0.140 (1.77) | 0.109 (1.20) | 0.078 (0.84) | 1.12 | 0.49 |
| *bm* | 9.36 | 0.174 (1.60) | 0.054 (0.49) | -0.006 (-0.06) | 0.68 | 0.20 |
| *roe* | 7.32 | 0.053 (0.81) | 0.158 (2.01) | 0.159 (1.94) | 0.51 | 0.13 |
| *ag* | 10.08 | 0.210 (2.99) | 0.214 (2.47) | 0.193 (2.23) | 1.40 | 0.38 |
| *mom12* | 7.68 | -0.049 (-0.54) | 0.043 (0.34) | 0.018 (0.13) | 1.62 | 0.42 |
| *rev36* | 9.60 | 0.127 (1.38) | 0.100 (0.91) | 0.054 (0.47) | 1.36 | 0.37 |
| *tvol* | 7.32 | 0.169 (1.57) | 0.203 (2.23) | 0.191 (2.04) | 0.82 | 0.21 |
| *accr* | 6.96 | -0.027 (-0.36) | 0.123 (1.26) | 0.078 (0.80) | 1.10 | 0.29 |
| Critical value: | | 0.213 | 2.48 | 2.44 | | |
| *B.* Equal weighting | | | | | | |
| *mve* | 11.04 | 0.133 (1.70) | 0.129 (1.56) | 0.100 (1.16) | 1.14 | 0.54 |
| *bm* | 10.92 | **0.238 (2.36)** | 0.134 (1.41) | 0.063 (0.70) | 0.72 | 0.29 |
| *roe* | 10.44 | **0.171 (2.56)** | **0.176 (2.39)** | 0.155 (2.14) | 0.77 | 0.27 |
| *ag* | 10.08 | 0.147 (1.87) | 0.126 (1.46) | 0.110 (1.25) | 1.38 | 0.48 |
| *mom12* | 10.20 | 0.068 (0.92) | 0.102 (0.87) | 0.101 (0.79) | 1.61 | 0.49 |
| *rev36* | 9.00 | 0.043 (0.49) | -0.029 (-0.28) | -0.080 (-0.75) | 1.29 | 0.46 |
| *tvol* | 9.12 | **0.318 (2.70)** | **0.268 (2.83)** | 0.227 (2.40) | 0.79 | 0.27 |
| *accr* | 9.48 | 0.067 (0.93) | 0.088 (0.91) | 0.047 (0.49) | 1.28 | 0.43 |
| Critical value: | | 0.161 | 2.25 | 2.41 | | |
| *C.* Score weighting | | | | | | |
| *mve* | 10.92 | 0.122 (1.62) | 0.146 (1.69) | 0.118 (1.29) | 1.37 | 0.68 |
| *bm* | 11.64 | **0.269 (2.48)** | 0.179 (1.68) | 0.114 (1.10) | 0.80 | 0.31 |
| *roe* | 10.32 | 0.149 (2.08) | 0.164 (1.90) | 0.156 (1.83) | 0.79 | 0.27 |
| *ag* | 9.96 | 0.121 (1.45) | 0.096 (0.98) | 0.078 (0.78) | 1.49 | 0.53 |
| *mom12* | 10.68 | 0.053 (0.65) | 0.122 (0.90) | 0.126 (0.84) | 1.66 | 0.50 |
| *rev36* | 9.00 | 0.011 (0.12) | -0.053 (-0.45) | -0.105 (-0.86) | 1.36 | 0.48 |
| *tvol* | 8.76 | **0.306 (2.33)** | **0.269 (2.46)** | 0.231 (2.11) | 0.87 | 0.29 |
| *accr* | 9.24 | 0.032 (0.41) | 0.070 (0.64) | 0.026 (0.22) | 1.36 | 0.46 |
| Critical value: | | 0.173 | 2.30 | 2.38 | | |

*Note*: The market portfolio generates annualized mean excess return of 6.72% and Sharpe ratio of 0.422. The number in parentheses is the *t*-statistic with Newey-West standard error. The *t*-statistic for Sharpe ratio difference is based on delta method (see Section 3.1 of Ledoit and Wolf [2008]). The critical value is calculated by performing multiple testing with FDP exceedance control at levels $\xi$=0.05 and $\delta$=0.05.

Exhibit 3 displays the long-only portfolio performance with respect to the top 20% stocks, based on single-factor, from the 500 largest stocks at the end of June. All of the

long-only single-factor portfolios have greater annualized mean excess return than that of the market portfolio, or 6.72%, and almost all of them have greater Sharpe ratios than that of the market portfolio, with the exceptions being high momentum and low accrual value-weighted portfolios. Nevertheless, only a few of them deliver significant outperformance. Panel A shows that none of the single-factor value-weighted portfolios significantly beat the market portfolio under the multiple testing framework. The low asset growth portfolio has the largest Sharpe ratio difference, 4-factor alpha, and 4-factor+*LIQ* alpha. While their corresponding individual *t*-ratios surpass the conventional threshold based on individual testing, they are merely near the critical value based on the multiple testing. The same is true for the alpha of the low volatility portfolio, the *t*-ratio for the 4-factor (4-factor+*LIQ*) alpha is 2.23 (2.04), which is less than the required hurdle rate 2.48 (2.44) to be significant.

In general, the portfolio performance is improved if we switch to equal or score weighting scheme. The portfolio turnover and transaction cost increase without entirely eliminating its profitability. For instance, the value-weighted portfolio *roe* has its mean excess return of 7.19% after transaction cost, while the equal-weighted portfolio *roe* delivers ex-transaction cost mean excess return of 10.17% along with significant Sharpe ratio difference and 4-factor alpha. The score weighting also significantly improves the Sharpe ratios of portfolios *bm* and *tvol* over the value weighting counterparts.

The evaluation of single-factor portfolios in Exhibit 3 reveals that exploiting the low risk anomaly with equal or score weighting scheme consistently beats the market portfolio in terms of Sharpe ratio. Its outperformance is also robust to risk adjustment with 4-factor model. However, the addition of illiquidity risk factor *LIQ* helps explain the abnormal return. Taking multiple testing bias into consideration, the 4-Factor+*LIQ* alpha of low risk portfolio is not significant, despite that the *t*-ratio of the alpha for equal-weighted portfolio *tvol* is 2.40, which is quite close to the critical value 2.41.

Exhibit 4 reports the multi-factor portfolios that deliver significant 4-factor+*LIQ* alpha. As suggested by the evaluation of single-factor portfolios, the 4-factor+*LIQ* risk factor model poses a greater challenge to find significant abnormal returns. To save space, we omit the results regarding the 4-factor alpha and the portfolio turnover. The multiple testing with 4-factor alpha would reject more than 4-factor+*LIQ* alpha, but the numbers of

rejections are much smaller than the ones that are based on Sharpe ratio. There are more than one hundred portfolios for each weighting schemes that could beat the market portfolio in terms of Sharpe ratio, thus we do not list all of them here.

Exhibit 4. Long-Only Multi-Factor Portfolios with Significant 4-factor+*LIQ* Alphas.

| | Excess Return | Sharpe Ratio Difference | 4-Factor+*LIQ* Alpha | Transaction Cost |
|---|---|---|---|---|
| *A.* Value weighting | | | | |
| *mve*\|*bm*\|*roe*\|*ag*\|*tvol* | 12.12 | 0.438 (3.81) | 0.345 (3.29) | 0.36 |
| Critical value (No. of significant portfolios) | | 0.2 (117) | 3.24 (1) | |
| *B.* Equal weighting | | | | |
| *roe*\|*ag*\|*tvol* | 10.80 | 0.360 (4.01) | 0.296 (4.01) | 0.33 |
| *mve*\|*roe*\|*ag*\|*tvol* | 11.64 | 0.361 (3.74) | 0.286 (3.61) | 0.43 |
| *roe*\|*ag*\|*tvol*\|*accr* | 10.92 | 0.336 (3.85) | 0.274 (3.33) | 0.36 |
| *mve*\|*roe*\|*tvol* | 11.40 | 0.325 (3.38) | 0.270 (3.13) | 0.39 |
| *roe*\|*tvol* | 9.96 | 0.304 (3.49) | 0.251 (3.18) | 0.27 |
| *roe*\|*ag*\|*mom12*\|*tvol* | 10.80 | 0.342 (4.27) | 0.244 (3.01) | 0.36 |
| *mve*\|*roe*\|*ag*\|*tvol*\|*accr* | 11.16 | 0.308 (3.45) | 0.229 (2.99) | 0.44 |
| Critical value (No. of significant portfolios) | | 0.137 (226) | 2.95 (7) | |
| *C.* Score weighting | | | | |
| *roe*\|*ag*\|*tvol* | 10.68 | 0.368 (3.86) | 0.293 (3.75) | 0.35 |
| *roe*\|*tvol* | 10.20 | 0.331 (3.57) | 0.289 (3.48) | 0.28 |
| *mve*\|*roe*\|*ag*\|*tvol* | 11.64 | 0.353 (3.41) | 0.282 (3.20) | 0.48 |
| *mve*\|*roe*\|*tvol* | 11.52 | 0.342 (3.43) | 0.271 (3.13) | 0.43 |
| *roe*\|*ag*\|*tvol*\|*accr* | 10.56 | 0.319 (3.60) | 0.270 (3.35) | 0.38 |
| *roe*\|*ag*\|*mom12*\|*tvol* | 11.16 | 0.357 (4.24) | 0.259 (3.06) | 0.39 |
| *roe*\|*tvol*\|*accr* | 10.20 | 0.294 (3.59) | 0.243 (3.04) | 0.33 |
| *tvol*\|*accr* | 9.72 | 0.324 (3.53) | 0.240 (3.00) | 0.40 |
| Critical value (No. of significant portfolios) | | 0.167 (175) | 2.98 (8) | |

*Note*: See the note in Exhibit 3.

After the adjustments for illiquidity risk and multiple testing bias, we are able to identify 1, 7, and 8 significantly outperforming multi-factor portfolios for value, equal, and score weighting schemes, respectively. Although the multi-factor portfolios do not always dominate the single-factor ones in terms of mean excess returns, most of the Sharpe ratios are improved. For example, combining *roe* and *tvol* with score weighting provides mean excess returns of 10.2%, which lies between the single-factor portfolios *roe* (10.32%) and

*tvol* (8.76%), but its Sharpe ratio difference and 4-factor+*LIQ* alpha are greater than both single-factor portfolios. Moreover, the *t*-statistic for the 4-factor+*LIQ* alpha convincingly lies above the critical value 2.98.

The factor *bm* plays no role in achieving significant 4-factor+LIQ alpha. This is mainly due to the fact that its abnormal return is captured by the risk factor HML. However, discarding the linear risk factor models as the performance evaluation tool, *bm* could be used to construct highly profitable multi-factor portfolios. For example, the Sharpe ratio of equal-weighted portfolio *bm*|*roe*|*tvol* surpasses the market portfolio's by a margin of 0.354, which is the third largest magnitude. The diversification benefit to achieve greater Sharpe ratio by combining the best three individual factors follows the Fundamental Law of Active Investment.[9] Nonetheless, without relying on the value factor exposure, there are other variants of multi-factor portfolios, such as *roe*|*ag*|*tvol*, *mve*|*roe*|*tvol*, or *roe*|*ag*|*tvol*|*accr*, for the investors to obtain the diversification benefit. The overall results suggest that the small-cap and value factors are not mandatory to build superior multi-factor portfolios as implied by Arnott et al. [2013] or the multi-factor index in Exhibit 2.

We find that the number of factors to combine in achieving significant abnormal returns rarely exceed five. Among the value-weighted portfolios, the only outperforming combination that employs five factors takes *mve*, *bm*, *roe*, *ag*, and *tvol*. The value-weighted portfolio *mve*|*bm*|*roe*|*ag*|*tvol* has the greatest mean excess returns (12.12%), Sharpe ratio difference (0.438), and 4-Factor+*LIQ* alpha (0.345), even compared to all of the equal-weighted and score-weighted portfolios. While both equal and score weighting schemes are known to be more efficient, their largest multi-factor portfolio mean excess returns, Sharpe ratio difference, and 4-factor+*LIQ* alpha never exceed 12%, 0.4, and 0.3, respectively. However, as will be shown later, the superior performance of value-weighted portfolio *mve*|*bm*|*roe*|*ag*|*tvol* is not robust to tests with different sub-sample periods and alternative factor definitions.

[9] We thank the anonymous referee for pointing this out. The Fundamental Law (Grinold [1989]) and its generalized version (Clarke et al. [2002]) suggest that the outperformance of a portfolio can be attributed to the number of informative signals to implement, the strength of the signals, and the hurdles to implement the optimal portfolio.

The equal-weighted portfolio with combination of five factors fails to generate outperformance convincingly as well. Panel B of Exhibit 4 shows that the 4-factor+*LIQ* alpha of portfolio *mve*|*roe*|*ag*|*tvol*|*accr* is only marginally significant. Combining too many signals may inject excessive volatility into the portfolio. Note that while the annualized mean excess return of portfolio *mve*|*roe*|*ag*|*tvol*|*accr* is 11.16%, its Sharpe ratio difference is only 0.308 and it requires the greatest transaction cost.

Exhibit 5. Equal-Weighted Decile Portfolios from 1,000 Stocks Universe.

| | Excess Return | Sharpe Ratio Difference | 4-Factor+*LIQ* Alpha | Transaction Cost |
|---|---|---|---|---|
| *A.* Single-factor. | | | | |
| *mve* | 13.68 | 0.137 (1.45) | 0.120 (0.99) | 4.14 |
| *bm* | 13.92 | **0.248 (2.14)** | 0.107 (1.00) | 1.85 |
| *roe* | 12.12 | 0.199 (2.34) | 0.173 (1.91) | 1.25 |
| *tvol* | 9.24 | **0.375 (2.9)** | **0.254 (2.57)** | 0.43 |
| Critical value (No. of significant portfolios) | | 0.203 (2) | 2.32 (1) | |
| *B.* Multi-factor | | | | |
| *mve*\|*roe*\|*tvol* | 13.32 | 0.424 (3.66) | 0.361 (4.37) | 1.65 |
| *mve*\|*bm*\|*roe* | 16.44 | 0.326 (2.91) | 0.339 (3.70) | 2.50 |
| *mve*\|*roe*\|*tvol*\|*accr* | 13.68 | 0.397 (3.73) | 0.334 (4.01) | 1.70 |
| *mve*\|*roe*\|*mom12*\|*tvol* | 13.80 | 0.404 (4.19) | 0.331 (3.80) | 1.40 |
| *mve*\|*bm*\|*roe*\|*ag*\|*tvol* | 13.92 | 0.403 (3.42) | 0.324 (3.72) | 1.85 |
| *roe*\|*ag*\|*tvol* | 11.52 | 0.409 (4.21) | 0.317 (4.38) | 0.65 |
| *mve*\|*roe*\|*ag*\|*tvol* | 13.32 | 0.387 (3.44) | 0.313 (3.68) | 1.77 |
| *roe*\|*tvol* | 10.68 | 0.375 (3.90) | 0.312 (4.06) | 0.47 |
| Critical value (No. of significant portfolios) | | 0.195 (164) | 2.85 (28) | |

*Note*: The long-only decile portfolios are constructed by holding with equal weight the top 10% factor scores of the 1,000 largest stocks at the end of June each year.

One of the main reasons why the costs of trading the outperforming multi-factor portfolios are all below 0.5% is that we only select the stocks within the 500 largest market cap. Exhibit 5 illustrates how expanding the pool of stocks to select from could result in high transaction costs. We construct the decile portfolios by purchasing the top 10% based on the factor scores from the largest 1,000 stocks, thus they would hold around the same number of stocks as the top 20% of the 500 largest stocks strategy in each year. Panel A

shows four of the single-factor portfolios with the greatest Sharpe ratio. The cost of constructing small-cap equal-weighted portfolio could be as high as 4.14%, which offsets the mean excess return to approximately 9.5%.[10] When we trade larger stocks only, the ex-transaction cost mean excess return is 10.5%. While our estimates of transaction costs are imperfect, it sufficiently illustrates how severe the illiquidity problem could be in trading small-cap stocks.

The low volatility portfolio delivers the best risk-adjusted performance, while still has the lowest transaction cost. Panel B of Exhibit 5 lists the top eight outperforming multi-factor portfolios based on 4-factor+*LIQ* alpha. We find that constructing multi-factor portfolio with 1,000 stocks has the tendency to select the factor *mve* as a good predictor. Inclusion of *mve* always increases the transaction cost to above 1% annually. Nevertheless, the liquidity-adjusted abnormal returns are still significant for certain combinations of factors. For instance, the alpha for portfolio *mve*|*roe*|*tvol* is 4.3% per annum with *t*-statistics greater than the critical value 2.85. Excluding *mve* would lower the alpha to 3.7% per annum. However, we estimate the difference in transaction costs between *mve*|*roe*|*tvol* and *roe*|*tvol* is almost 1.2% per annum, which would offset the benefit of incorporating *mve*. The results so far show that combining factors to construct equal-weighted or score-weighted portfolio appears to be a promising method to obtain significant outperformance. The abnormal returns are free of illiquidity and multiple testing bias. The strategy also works when we add 500 less liquid stocks to select from. Next we analyze whether the outperformance of multi-factor strategy is robust to alternative definitions of factors and different subsample periods.

We follow Hsu et al. [2015], who recommend to perturb the definitions of the factors, to check the robustness of investment styles. Exhibit 6 reports the equal-weighted portfolios results when six measures of factors are changed. Some of the measures are shown to be superior factors in extant literature. For instance, George and Hwang [2004] study the 52-week high strategy as opposed to the momentum measure used in Jegadeesh and Titman [1993]; Novy-Marx [2013] suggests that gross profit scaled by total assets

[10] The value weighting scheme does not completely remedy the problem of high transaction cost. The estimated transaction cost for portfolio *mve* is 3.55%. However, the transaction costs for the other single-factor portfolios range from 0.21% to 0.86%.

provides a better measure of profitability. Panel B shows that *ey*, *gpa*, and *inv* have greater Sharpe ratio than *bm*, *roe*, and *ag*, respectively. It also suggests that *tvol* is slightly better than *beta* to exploit the low risk anomaly. The factor *gpa* is the only portfolio with significant 4-factor+*LIQ* alpha, this still holds when we consider value and score weighting schemes.

Exhibit 6. Portfolio Evaluation with Alternative Definitions of Factors.

*A.* Alternative definition of factors.

| Style | Factor | Definition |
|---|---|---|
| Value | *ey* | EBIT / market value of equity. |
| Momentum | *wh52* | Current stock price / 52-week high of stock price. |
| Long-term reversal | *rev60* | Cumulative stock returns over the past sixty months, excluding the recent year. |
| Risk | *beta* | CAPM beta from regression using the past 52 weekly returns. |
| Profitability | *gpa* | Gross profit / total asset. |
| Growth | *inv* | Capital expenditure / gross property, plant, and equipment. |

| | Excess Return | Sharpe Ratio Difference | 4-Factor+*LIQ* Alpha | Transaction Cost |
|---|---|---|---|---|
| *B.* Equal-weighted single-factor portfolios. | | | | |
| *ey* | 12.00 | **0.252 (2.48)** | 0.167 (1.93) | 0.37 |
| *gpa* | 10.80 | **0.216 (3.4)** | **0.315 (4.54)** | 0.23 |
| *inv* | 10.92 | **0.228 (2.44)** | 0.106 (1.15) | 0.38 |
| *wh52* | 9.24 | 0.173 (1.89) | 0.03 (0.26) | 0.51 |
| *rev60* | 10.44 | 0.13 (1.53) | 0.081 (0.76) | 0.45 |
| *beta* | 9.60 | **0.294 (2.81)** | 0.196 (1.85) | 0.50 |
| Critical value | | 0.138 | 2.31 | |
| *C.* Equal-weighted multi-factor portfolios. | | | | |
| *gpa\|beta* | 11.76 | 0.324 (3.54) | 0.407 (4.04) | 0.42 |
| *gpa\|beta\|accr* | 11.76 | 0.318 (3.74) | 0.376 (3.80) | 0.49 |
| *gpa\|rev60\|beta* | 11.52 | 0.294 (3.40) | 0.349 (3.26) | 0.44 |
| *gpa\|accr* | 12.36 | 0.237 (3.91) | 0.340 (4.55) | 0.43 |
| *gpa\|inv\|beta* | 10.80 | 0.357 (3.93) | 0.333 (3.57) | 0.44 |
| *ey\|gpa\|inv\|beta* | 11.52 | 0.402 (4.25) | 0.329 (3.96) | 0.42 |
| *gpa\|rev60* | 12.24 | 0.204 (2.71) | 0.324 (2.77) | 0.37 |
| *ey\|gpa\|beta* | 11.28 | 0.343 (3.41) | 0.317 (3.60) | 0.44 |
| Critical value (No. of significant portfolios) | | 0.126 (247) | 2.50 (93) | |

*Note*: The definition of six factors is altered to ensure the robustness of multi-factor investment strategies.

The significantly outperforming multi-factor portfolios with the new definitions of factors possess similar attributes to the results in Exhibit 4. The combination of high profitability, low growth, and low risk characteristics remains the superior one in this setting. Both portfolios *gpa*|*inv*|*beta* and *roe*|*ag*|*tvol* generate significant 4-factor+*LIQ* alpha and Sharpe ratio difference. The Sharpe ratios of portfolios *gpa*|*beta* and *roe*|*tvol* significantly surpass the market portfolio by the magnitudes of 0.324 and 0.304, respectively. In untabulated results, the total number of value-weighted (score-weighted) multi-factor portfolios with significant 4-factor+*LIQ* alpha is 5 (97). The number of significant multi-factor portfolios in this second set of factors increases substantially as expected since all single-factor portfolios perform significantly better except for *beta*. In sum, the superiority of multi-factor investment strategies is robust to the different measurement of factors.

The next step of our analyses is to examine the multi-factor portfolios performance with the most recent subsample. Recent study by Chordia et al. [2014] suggests that the post-1994 sample establishes the period of increased liquidity and trading activity in the U.S. stock market. They show that arbitrage activity in the market has reduced the abnormal returns of twelve factors in their study. We follow Chordia et al. [2014] to use the sample after June 1994 as the robustness test of multi-factor investment strategies in the period increased arbitrage activity. The June 1994 cut-off point splits our dataset into two subsamples of around 21 years.

Our results suggest that it indeed becomes harder to identify significantly profitable multi-factor portfolios. Using the first set of factors definition, we find no evidence in favor of multi-factor portfolios outperformance with equal or value weighting scheme. None of the 4-factor+*LIQ* alphas surpass the corresponding critical value. In Panel A of Exhibit 7, we list the only two of the score-weighted portfolios which have significant 4-factor+*LIQ* alphas. Of these two portfolios, *roe*|*mom12*|*tvol* fails to pass the multiple testing criteria in the pre-1994 period. Meanwhile, the portfolio *roe*|*tvol* has significant alpha, but its Sharpe ratio difference falls slightly below the respective critical value.

The second set of factors definition provides supports for multi-factor investment strategies. Using the value, equal, and score weighting schemes result in 3, 6, and 15

portfolios with significant 4-factor+*LIQ* alphas, respectively. Panel B of Exhibit 7 shows 8 of the outperforming score-weighted portfolios. Most of the portfolios deliver stable performance in both pre- and post-1994 sample periods, with the exception being portfolio *gpa|rev60|beta*.

Exhibit 7. Portfolio Evaluation in the Post-1994 Sample Period.

| | Excess Return | Sharpe Ratio Difference | 4-Factor+*LIQ* Alpha | Transaction Cost |
|---|---|---|---|---|
| *A.* Score-weighted portfolios with the first set of factor definition. | | | | |
| *roe\|mom12\|tvol** | 12 | 0.419 (3.03) | 0.369 (3.12) | 0.37 |
| *roe\|tvol* | 11.04 | 0.380 (2.21) | 0.358 (3.19) | 0.28 |
| Critical value (No. of significant portfolios) | | 0.402 (1) | 3.03 (2) | |
| *B.* Score-weighted portfolios with the second set of factor definition. | | | | |
| *gpa\|rev60\|beta** | 12.36 | 0.439 (2.73) | 0.449 (3.83) | 0.50 |
| *gpa\|beta\|accr* | 12.24 | 0.444 (2.82) | 0.447 (3.87) | 0.52 |
| *ey\|gpa\|inv\|beta* | 12.12 | 0.405 (2.33) | 0.420 (3.72) | 0.48 |
| *gpa\|inv\|beta* | 11.52 | 0.421 (2.46) | 0.418 (3.56) | 0.50 |
| *ey\|gpa\|beta* | 12.12 | 0.377 (2.06) | 0.408 (3.11) | 0.49 |
| *gpa\|beta* | 11.28 | 0.401 (2.36) | 0.402 (3.08) | 0.47 |
| *gpa\|accr* | 12.60 | 0.260 (2.50) | 0.400 (3.95) | 0.45 |
| *gpa\|rev60\|beta\|accr* | 12.00 | 0.383 (2.52) | 0.385 (3.78) | 0.53 |
| Critical value (No. of significant portfolios) | | 0.409 (5) | 2.92 (15) | |

*Note*: This exhibit shows some of the score-weighted portfolios with significant 4-factor+*LIQ* alpha in the post-1994 period. The sign * indicates that the multi-factor portfolio is only significant in the post-1994 but not in pre-1994 test period.

In the post-1994 sample period, there appear to be increases in the estimated critical values for Sharpe ratio. The required difference for Sharpe ratio to significantly beat the market portfolio is approximately 0.4 in the post-1994 era, while the critical values for Sharpe ratio difference in the pre-1994 are around 0.2. Recall that the critical value of multiple testing depends on the distribution of *k*-th largest statistics. This suggests that the variability of Sharpe ratios among the outperforming portfolios has been much larger, and may explain why it becomes harder to find significant multi-factor strategies in this period.

***Long-short portfolios.*** We examine if the multi-factor scores could generate significantly risk-adjusted long-short spreads, which may be of interests to other investors who do not face short-selling constraint. The short-leg portfolios consist of the stocks with

combined scores below the bottom 20% quantile. The score-weighted portfolios are constructed with the similar weighting formula in Equation (8) by replacing the upper quantile with the bottom quantile.

Exhibit 8. Long-Short Portfolios.

| Equal-weighting | | Score-weighting | |
|---|---|---|---|
| Top performing portfolios | 4-Factor+*LIQ* | Top performing portfolios | 4-Factor+*LIQ* |
| *A.* First set of factors (full sample). | | | |
| *mve*\|*roe*\|*ag*\|*tvol* | 0.449 (3.27) | *mve*\|*roe*\|*ag*\|*mom12*\|*tvol* | 0.478 (3.36) |
| *roe*\|*ag*\|*tvol*\|*accr* | 0.385 (3.14) | *roe*\|*ag*\|*mom12*\|*tvol* | 0.473 (3.24) |
| *mve*\|*roe*\|*ag*\|*tvol*\|*accr* | 0.378 (3.44) | *roe*\|*ag*\|*mom12*\|*tvol*\|*accr* | 0.454 (3.46) |
| | 2.99 (5) | | 3.0 (9) |
| *B.* Second set of factors (full sample). | | | |
| *ey*\|*gpa*\|*beta* | 0.472 (3.18) | *ey*\|*gpa*\|*beta* | 0.609 (3.54) |
| *gpa*\|*beta*\|*accr* | 0.460 (3.76) | *ey*\|*gpa*\|*inv*\|*beta*\|*accr* | 0.548 (4.35) |
| *gpa*\|*beta* | 0.459 (2.90) | *ey*\|*gpa*\|*beta*\|*accr* | 0.544 (4.05) |
| | 2.82 (38) | | 2.58 (64) |
| *C.* Second set of factors (post-1994 sample). | | | |
| *gpa*\|*beta* | 0.620 (3.03) | *ey*\|*gpa*\|*beta* | 0.712 (3.34) |
| *ey*\|*gpa*\|*beta* | 0.557 (3.14) | *gpa*\|*inv*\|*beta* | 0.689 (3.04) |
| *gpa*\|*beta*\|*accr* | 0.508 (3.40) | *gpa*\|*rev60*\|*beta* | 0.682 (3.04) |
| | 2.92 (5) | | 2.94 (6) |

*Note*: This table shows three portfolios with the greatest 4-factor+*LIQ* alphas. The last row of each panels reports the critical value along with the number of significant portfolios in the parentheses.

Exhibit 8 reports the top three significantly outperforming long-short portfolios based on the 4-factor+*LIQ* abnormal returns. None of the long-short value-weighted portfolios deliver significant alphas. Panels A and B show the test results using the full sample. The results suggest that multi-factor investment strategies also work for constructing long-short portfolios. However, the long-short portfolios may employ slightly different combinations from the long-only strategies to achieve significant risk-adjusted returns. For instances, in contrast to the long-only portfolios universe, *roe*|*tvol* and *roe*|*ag*|*tvol* are not among the score-weighted portfolios with significant 4-factor+*LIQ* alphas. For the long-short portfolios, adding the factor *mom12* is needed to achieve significant risk-adjusted spreads.

In the post-1994 sample period, we fail to identify any long-short portfolios with significant 4-factor+*LIQ* alphas using the first set of factors definition. Applying the recent

findings of new factors, such as *gpa* by Novy-Marx [2013], improves the performance of long-short portfolios. The combination *ey*|*gpa*|*beta* generates significant risk-adjusted spread of 0.712 and 0.557 based on score and equal weighting portfolio methods, respectively. We also note that, in the results not shown here, the standalone single-factor portfolios *gpa* and *beta* produce insignificant risk-adjusted spreads in the post-1994 period, but they are still crucial factors in constructing significantly outperforming multi-factor portfolios.

## Conclusion

In this study, we evaluate the multi-factor portfolio performances with multiple testing methodology to adjust for the data mining bias in statistical inference. The overall results suggest that there exists a potential benefit for investors to increase their portfolio exposure to multi-factor investment strategies. However, the number of significantly outperforming portfolios becomes smaller in the past two decades. One of the profitable investing styles is a combination of high profitability and low volatility firm characteristics, in which the illiquidity risk-adjusted returns of such strategies are significant based on high standard. Moreover, they are robust to alternative definitions of factors and different subsamples.

Consistent with the literature, the value-weighting scheme is not the best way to exploit profitability of factor investing. While switching to equal-weighting or score-weighting increases the turnover, the estimated transaction costs for most multi-factor portfolios are still below 0.5%. The restriction to select only from the 500 largest market-cap stocks is particularly crucial to reduce the transaction costs. Nevertheless, our empirical analyses suggest that the selection from highly liquid stocks does not hinder the outperformance of multi-factor strategies.

## Appendix A

This appendix presents the Step-SPA($k$) and FDP-SPA procedure to test the multiple inequalities

$$H^i: \theta_i \leq 0, \qquad i = 1, \dots, M.$$

Let $max(A, k)$ and $1(\cdot)$ denote the $k$-th largest value of vector $A$ and the indicator function, respectively; $X_i$ denotes the vector of excess returns of portfolio $i$; and $Y$ denotes the matrix of risk factor portfolios or the benchmark portfolio. There are two versions of test statistics that could be used: non-studentized and studentized. In the portfolio evaluation using alphas, we use the studentized test statistics, i.e. the $t$-ratio, so that they are comparable on the same scale. However, we use non-studentized test statistics when examining the Sharpe ratio differences.

The recentering estimator $\hat{\theta}_i^-$ is used to enhance the power of the test. Hansen [2005] shows that only the set of models or portfolios with $\theta_i = 0$ would affect the critical value of the $k$-th largest $\theta$. Furthermore, the statistical power of multiple inequality testing could be substantially reduced if too many "irrelevant" inferior models are included. That is, if we can determine which $\theta_i < 0$, then we increase the power while maintaining the control of $k$-FWER. Hansen [2005] recommends the use of threshold $-\hat{\sigma}_i\sqrt{2\log\log T}$, which is based on law of iterated logarithm, for $\sqrt{T}\hat{\theta}_i$.

Step-SPA($k$) algorithm with level $\delta$

**procedure** stepSPA($\{X_1, \dots, X_M, Y\}, \delta, k$)
create vector STAT of size $M$
**for** $i \in \{1, \dots, M\}$ **do**
    calculate the parameter of interest $\hat{\theta}_i$ and its standard error $\hat{\sigma}_i$
    $\hat{\theta}_i^- = \hat{\theta}_i \times 1(\sqrt{T}\hat{\theta}_i \leq -\hat{\sigma}_i\sqrt{2\log\log T})$
    STAT[$i$] = $\sqrt{T}\hat{\theta}_i$ or $\sqrt{T}\hat{\theta}_i/\hat{\sigma}_i$ ◆ non-studentized or studentized test statistics
**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**

compute $\hat{\theta}_i^s$ with the bootstrap sample

$\mathrm{X}[i,s] = \sqrt{T}(\hat{\theta}_i^s - \hat{\theta}_i + \hat{\theta}_i^-)$ or $\sqrt{T}(\hat{\theta}_i^s - \hat{\theta}_i + \hat{\theta}_i^-)\,/\hat{\sigma}_i$ ◆ non-studentized or studentized test statistics

**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 sort_index

NUM_REJECT = 0

NUM_REJECT1 = -1

create vector KMAX of size $B$

**while** NUM_REJECT > NUM_REJECT1 **do** ◆ The procedure will stop when there is no further rejection

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, $round(\delta \times B)$ )

**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 $\alpha$ and $\xi$

**procedure** FDP_SPA($\{X_1, \dots, X_M, Y\}, \delta, \xi$)

$k$ = 1

CRITICAL_VALUE = stepSPA($\{X_1, \dots, X_M, Y\}, \delta, k$)

NUM_REJECT = $sum$(1(STAT > CRITICAL_VALUE))

**while** NUM_REJECT < $k/\xi - 1$ **do**

6 $k = k + 1$

7 CRITICAL_VALUE = stepSPA($\{X_1, \ldots, X_M, Y\}, \delta, k$)

8 NUM_REJECT = $sum$(1(STAT > CRITICAL_VALUE))

9 **end while**

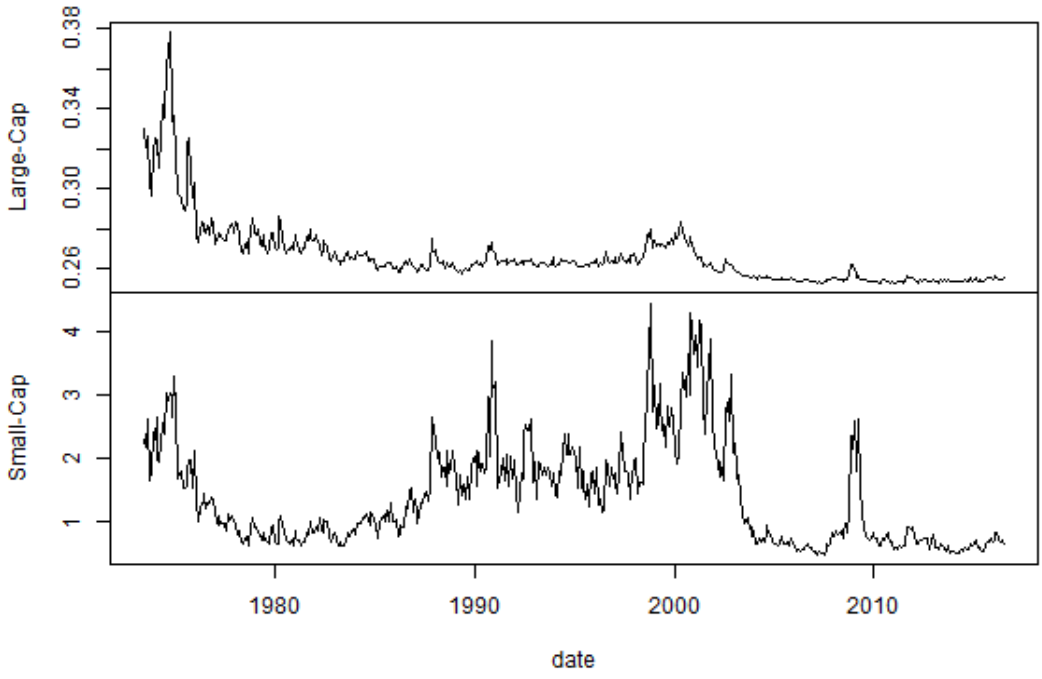
10 Output: CRITICAL_VALUE

11 **end procedure**

## Appendix B

To demonstrate the difference in estimated transaction costs between trading large-cap and trading small-cap stocks sampled in this study, we provide time-series plots for the median estimates for both categories in Exhibit B1. We follow Fama and French [1993] to split the stocks into large-cap and small-cap with the NYSE median of market capitalization at the end of each month. For the large-cap stocks, we document that the transaction costs are trending downward in the past few decades. This result is consistent with the literature, see e.g. Chorida et al. [2014]. The median of the transaction cost is estimated to be between 0.25 and 0.3 for the large-cap stocks since 1976. The transaction costs are considerably greater for small-cap stocks, especially in the times of stock market downturn. In the period of crises, the high price impact would cause the spike in the costs of trading small-cap stocks.

Exhibit B1. The Estimated Transaction Cost 1973-2016.




*Note*: The figures depict the monthly median transaction cost for large-cap and small-cap stocks universe from June 1973 to June 2016. At the end of each month, all of the stocks are classified into large-cap and small-cap based on the NYSE median of market capitalization.

## Reference


Acharya, V. V. and Pedersen, L. H. (2005). "Asset pricing with liquidity risk." *Journal of Financial Economics*, **77**, 375-410.

Amenc, N., Ducoulombier, F., Goltz, F., Lodh A., and Sivasubramanian (2016). "Diversified or concentrated factor tilts?" *Journal of Portfolio Management*, **42**, 64-76.

Amihud, Y. (2002). "Illiquidity and stock returns: Cross-section and time-series effects." *Journal of Financial Markets*, **5**, 31-56.

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.

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). "Value and momentum everywhere." *Journal of Finance*, **68**, 929-985.

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.

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.

Carhart, M. (1997). "On persistence in mutual fund performance." *Journal of Finance*, **52**, 57-82.

Chordia, T., Subrahmanyam, A., and Tong, Q. (2014). "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?" *Journal of Accounting and Economics*, **58**, 41-58.

Clarke, R., de Silva, H., and Thorley, S. (2002). "Portfolio constraints and the fundamental law of active management" *Financial Analysts Journal*, **58**, 48-66.

Cochrane, J. H. (2011). "Presidential address: Discount Rates." *Journal of Finance*, **66**, 1047-1108.

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

Delattre, S. and Roquain, E. (2015). "New procedures controlling the false discovery proportion via Romano-Wolf's heuristic." *Annals of Statistics*, **43**, 1141-1177.

Fama, E. F. and French, K. R. (1992). "The cross-section of expected stock returns." *Journal of Finance*, **47**, 427-465.

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

George, T. J. and Hwang, C.-Y. (2004). "The 52-week high and momentum investing." *Journal of Finance*, **59**, 2145-2176.

Greenblatt, J. (2010) *The Little Book That Still Beats the Market*. John Wiley & Sons.

Grinold, R. C. (1989). "The fundamental law of active management" *Journal of Portfolio Management*, **15**, 30-38.

Goyenko, R. Y., Holden, C. W. and Trzcinka, C. A. (2009). "Do liquidity measures measure liquidity?" *Journal of Financial Economics*, **92**, 153-181.

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.

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. and Titman, S. (1993). "Returns to buying winners and selling losers: The implications for stock market efficiency." *Journal of Finance*, **48**, 65-91.

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.

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

Malkiel, B. G. (2014). "Is smart beta really smart?" *Journal of Portfolio Management*, **40**, 127-134.

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. (2013). "The other side of value: The gross profitability premium." *Journal of Financial Economics*, **108**, 1-28.

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

Pastor, L. and Stambaugh, R. F. (2003). "Liquidity risk and expected stock returns." *Journal of Political Economy*, **111**, 642-685.

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

Stambaugh, R. F. and Yu, Y. (2017). “Mispricing factors.” *Review of Financial Studies*, **30**, 1270-1315.

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