The Journal of Portfolio Management

VOLUME 43, NUMBER 2 www.iijpm.com WINTER 2017

Quantifying Backtest Overfitting in Alternative Beta Strategies

ANTTI SUHONEN, MATTHIAS LENNKH, AND FABRICE PEREZ

IIJ Institutional Investor Journals

The Voices *of* Influence | iijournals.com

Quantifying Backtest Overfitting in Alternative Beta Strategies

ANTTI SUHONEN, MATTHIAS LENNKH, AND FABRICE PEREZ

ANTTI SUHONEN
is a professor of finance at the Aalto University School of Business in Helsinki, Finland.
antti.suhonen@aalto.fi

MATTHIAS LENNKH
is a director at Clear Alpha Limited in London, U.K.
matthias.lennkh@clearalpha.com

FABRICE PEREZ
is a director at Clear Alpha Limited in London, U.K.
fabrice.perez@clearalpha.com

Alternative beta strategies seeking to extract factor returns beyond directional market beta, or to exploit apparent market anomalies, have been one of the major new investment trends of recent years. Analysis by Morningstar (Bioy et al. [2015]) reports collective assets under management (AUM) of USD 497 billion across 844 individual "strategic beta" exchange-traded products worldwide. Today, the "alternative beta"[1] universe extends far beyond alternative weighting schemes to market capitalization in equities and represents a diverse range of strategies embodying systematic trading algorithms, use of derivatives, and leverage across all liquid asset classes. Recent interest from large institutional investors has been an important contributor to the growth of the investment style. In a related development, global investment banks have emerged alongside asset managers as providers of alternative beta strategies over the past decade. Although no official gauge of the size of the investment-bank-promoted market exists, typical estimates by industry sources are in the region of USD 100 to 200 billion.[2]

The identification, development, and implementation of alternative beta strategies are inherently quantitative exercises that often rely heavily on the analysis of historical data. Consequently, such strategies are susceptible to biases arising from data mining, multiple testing, and selective reporting that have been the focus of a growing volume of recent research (e.g., Bailey et al. [2014], Harvey et al. [2014], and Harvey and Liu [2015]).

Our objectives in this article are as follows. First, we will present an overview of the investment-bank-sponsored alternative beta market that is growing in significance. To our knowledge, no previous studies examining the performance and risk characteristics of the broad market of alternative beta strategies sponsored by investment banks across asset classes exist. Second, we will examine the realized returns and risks of the strategies and, specifically, the persistence of risk-adjusted returns after the "live" date—that is, when the strategies are launched in the market with a final and published investment algorithm. This allows us to quantify the possible biases in strategy construction highlighted in prior research. Third, we'll investigate the robustness of the factor exposures of four selected strategy families (equity value, equity volatility, fixed income curve, and FX carry) in the backtest and live periods. We use the Fama–French–Carhart (Fama and French [1992]; Carhart [1997]) four-factor model for the equity value and volatility strategies and the Fung and Hsieh [2001, 2004] primitive trend-following strategy returns and market benchmarks as a starting point for analyzing the other two strategies.

We use a unique, proprietary data set composed of the daily returns of 215 alternative beta strategies across five asset classes, eleven identifiable strategy groups, and fifteen sponsor investment banks. It lends itself well to a natural experiment of backtested biases for the following reasons:

1. Investment-bank-sponsored alternative beta strategies are formulaic and nondiscretionary strategies that follow set investment rules (asset selection, rebalancing, execution, risk management, etc.) that typically cannot be changed by the index sponsor after the live date.
2. Upon the release of a new alternative beta strategy, the sponsoring investment bank will provide investors with a backtested time series illustrating the hypothetical past performance of the proposed strategy. Given the immutability of the strategy rules, there is some justification for arguing that this past performance would have some value in assessing the future behavior of the strategy.
3. Our sample contains the daily backtested (pre-live date) performance of each strategy and the daily live (post-live date) performance, as well as the exact live date as reported by the strategy sponsor. The average length of the backtest periods is 10.7 years and that of the live periods is 4.6 years.

Furthermore, the strategies in our database have been selected for commercial promotion by the sponsoring investment banks, presumably following a rigorous analysis of the practical feasibility of the strategy. In other words, the database should be reasonably free of biases arising from strategies based on factors or anomalies that might exhibit attractive risk-adjusted returns in a theoretical or historical context but suffer from poor market liquidity, excessive transaction costs, or other structural impediments to real-life implementation and investment.

Our results strongly support the existing literature on selection and publication biases and backtest overfitting. We find a median Sharpe ratio (SR) of 1.20 across the 215 alternative beta strategies during their respective backtest periods, compared to 0.31 during live performance. In the backtest period, 95% of the strategies have an SR that is positive and statistically significant at the 10% level (i.e., p-value < 0.1), and 80% are significant at the 1% level according to a standard t-test, dropping to just over a quarter being significant at the 10% level after going live. A decline is evident across all asset classes and strategy groups, but it appears most pronounced in the FX asset class and in event-driven, value, and carry strategies. Multi-asset, fixed income, and volatility strategies, as well as trend-following strategy groups, exhibit a relatively smaller proportional decline in risk-adjusted returns between the backtested and live periods.

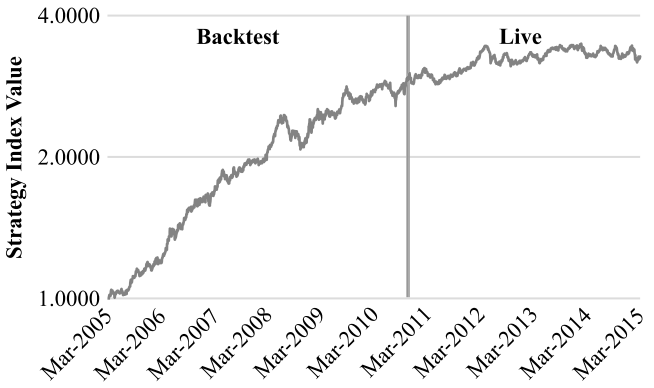
We also find a link between alternative beta strategy complexity and the deterioration of live versus backtested performance as more complex strategies—that is, those including more trading rules, filters, and parameters—appear more likely to suffer from backtest overfitting. Our results indicate that the SR "haircuts," or the percentage reduction of the SR in live versus backtested time series of the most complex strategies, are over 30 percentage points higher than those of the simplest strategies, after controlling for strategy, asset class, and vintage fixed effects.

Exhibit 1 illustrates our key findings. The figure shows the backtested and live performance (excess return index) of one alternative beta strategy in the sample over a 10-year period, with the vertical line in October 2010 marking the live date of the strategy. The strategy is representative of the median in our sample, with a backtested SR of 1.42 and live SR of 0.36, or a realized SR haircut of 75%.

Our analysis of the selected four strategy families illustrates heterogeneity in strategy robustness. For the equity value family, we find a highly significant loading on the value factor in the backtest, which is erased in

EXHIBIT 1

Illustration of Strategy Performance in Backtest and Live Periods (log scale)



the live period. In contrast, the equity value strategies have a positive relationship with market returns and the size and momentum factors during the live period. The equity volatility, fixed income curve, and FX carry strategy families exhibit more consistent exposures to their respective return drivers (implied and realized equity volatility, changes in bond yields, and returns to a primitive currency carry model, respectively). All of the four strategy families examined in detail show a statistically significant positive alpha in factor regressions during the backtest period, whereas in the live period the alphas are generally reduced by half. The live alphas remain positive and statistically significant for fixed income curve, equity volatility, and, at the margin, the FX carry strategy. The equity volatility alpha turns insignificant when we introduce to the regression a tradable instrument (iPath S&P 500 VIX Short-Term Futures ETN) that captures the structural contango in the equity volatility term structure.

Our investigation builds on recent practical and academic debate on the inherent biases in research and investment strategy design.[3] McLean and Pontiff [2016] review the post-publication performance of 97 variables that academic research has shown to predict cross-sectional stock returns. The authors find that the returns are 26% lower out of sample and 58% lower post-publication, indicating both a data mining effect (evidenced by the lower out-of-sample performance) and a crowding effect (investors learning about a mispricing from academic publications).[4]

The impact of inherent biases arising from data mining, multiple testing, and the tendency to publish only positive results has been the focus of a number of recent research papers. Harvey et al. [2014] analyze 315 factors reported in published finance research seeking to explain the cross section of stock returns. The authors conclude that many of the seemingly significant factors reported in studies do not meet the significance criteria when adjusted for the number of previous trials. Harvey and Liu [2015] discuss the "common practice" of applying a 50% haircut on the Sharpe ratios of new strategies and argue that in fact a nonlinear adjustment accounting for the number of previous trials should be applied. Harvey and Liu [2014] suggest methods to account for multiple testing and consider a practical example using published quantitative trading strategies, whereas Bailey and Lopez de Prado [2014] propose a correction for multiple testing and non-normal distribution of strategy returns based on additional information of the strategy development process and statistical properties of strategy returns.

Bailey et al. [2014] propose a model for the minimum track record needed when multiple tests or strategies are evaluated. The authors discuss the practice of model overfitting in the financial services industry and note that there is likely to be a positive link between the complexity of a model and the probability of overfitting, because the increase in a model's parameters make the fitting to historical data easier. Investment strategies are susceptible to overfitting at different levels. First, the core investment algorithm may be designed to perform well (or less badly) in sample at known times of historical market stress—for example, during the 2008–2009 financial crisis. Second, the specific allocation, risk management, rebalancing, and execution rules are often optimized in sample. Third, parametric optimization methods rely on assumptions made about the statistical properties of underlying data. Finally, the strategy development process does not necessarily control for the number of trials attempted before the ultimate configuration was selected—leading to a false positive due to multiple testing. Hence, reported disappearances of, for example, seasonal effects may not be the result of an anomaly being arbitraged away, but rather the result of a "false positive" finding in the first place.

In other related research, Amenc et al. [2015a] discuss the concept of robustness in the context of equity "smart beta" strategies. The authors use two separate definitions of robustness: *relative robustness*, a strategy's ability to offer similar performance in similar market conditions, and *absolute robustness*, a strategy's capacity to outperform regardless of prevailing market conditions. Their paper highlights the risk of unintended factor tilts in the strategies, leading to a lack of relative robustness.

This study makes several contributions to the existing literature. We use a unique database that lends itself to a natural experiment in measuring the biases in economic testing. We extend the investigation beyond equity strategies, and offer empirical support to findings reported in prior research and by industry practitioners. Our results also have implications on wider research of investment management, banking, and structured products. Finally, we believe the results have important practical significance for institutional investment policy, strategy development and selection, and portfolio construction. We refer to a recent survey of

institutional investors in alternative equity beta (Amenc et al. [2015b]) in which respondents list the lack of access to data, lack of transparency on strategy methodology, and difficulty of evaluating specific risks and factor tilts as the key challenges for investors.

ALTERNATIVE BETA STRATEGIES[5]

Alternative beta investing has attracted significant interest among the institutional investor community in recent years, thanks in no small part to a number of influential publications combining theory and practice in the field. Ang et al. [2009] and Ang [2014] discuss the foundations and practice of factor investing in an institutional investment setting, and Ilmanen [2011] provides a comprehensive overview of the evidence, sources, and possible explanations behind premia observed across asset classes, risk factors, and investment styles.

Alternative beta strategies seek to identify, isolate, extract, and monetize premia and market anomalies in an investible format. The basic building blocks of many of the strategies are well documented in research, and the strategies have been widely used by practitioners, particularly hedge funds, in the asset management industry over several decades. In addition to providing access to the sources of factor risk premia, or excess returns arising from apparent market anomalies, alternative beta products may also provide investors with diversification across a range of factors or provide time-varying exposure to factors according to systematic algorithms. As such, one can argue that many of the strategy families meet at least some of the factor criteria of Ang et al. [2009], for example:

1. They are justified by academic research.
2. They have exhibited significant premia that are expected to persist in the future.
3. They have historical return data available for "bad times."
4. They are implementable using liquid, traded instruments.

Alternative beta strategies fundamentally rely on the accurate identification, robustness, and persistence of the underlying risk factors, patterns, and anomalies (essentially, criteria 2 and 3). Therefore, we should expect the performance of alternative beta strategies to reflect the risk factor they are designed to capture, such as the difference between implied and realized volatility, or the steepness of the short end of the interest rate swap curve in a given market. In other words, we would expect alternative beta strategies to experience gains and losses in well-identified circumstances. If a strategy does not reflect the increase or decrease of an identified risk factor, there may be other effects at work that require investigation—for instance, the existence of a market anomaly that does not have a risk-based explanation or, indeed, of a bias in the way the in-sample performance has been achieved.

The global alternative beta market is estimated to be several hundred billion dollars in size. Long-only "smart beta" products linked to equity markets represent the majority of invested assets, but recent years have also seen a growth in more complex strategies involving long and short positions and leverage across various asset classes. The providers of investible products in alternative beta encompass not only asset managers, but also global investment banks. Banks started the development of alternative beta strategies over a decade ago, but it was only during and after the financial crisis of 2008–2009 that the market has witnessed significant growth. Alternative beta is an interesting phenomenon that straddles the traditional capital markets and brokerage/execution roles of investment banks, as well as investment research, portfolio management, and other functions that have historically been the domain of asset managers.

Some industry sources trace the origins of alternative beta strategies sponsored by investment banks to the structured products markets, with banks looking to achieve product differentiation by offering products linked to in-house strategies alongside broadly traded benchmark indices. Although the target clientele for structured products is predominantly individual investors, either via banking and wealth management channels or through life insurance products, institutional investors have become a major source of alternative beta demand in recent years. In the 2014 survey by Rabier and Suhonen, pension funds, insurance companies, and sovereign wealth funds were listed as the most important investor types by the participating investment banks.

Alternative beta products created by investment banks are formulaic, prescribed, and nondiscretionary investment strategies that aim to give investors exposure to the premia from specific risk factors and investment styles (e.g., value, momentum, carry, or interest rate term premia) or returns arising from anomalies or

statistical biases in financial markets (e.g., turn-of-the-month effects in equity markets, commodity congestion arbitrage, or mean reversion in equity indices). The market also comprises strategies based on research inputs from a variety of systems and models, as well as broad asset allocation systems driven by technical or macroeconomic inputs.

The product development process for alternative beta products varies across the different investment banks but usually involves collaboration across trading, structuring, and research functions. Once a new, interesting strategy has been identified, developed, and selected for commercial production, the bank will publish an index representing the hypothetical past returns of the strategy to its clients and the public domain via market information services such as Bloomberg. Such publication usually coincides with the start of a marketing campaign toward prospective investors. Crucially, for the purposes of this article, the set of rules, parameters, and algorithm of the strategy index are fully described in a methodology document that will not change after initial publication, apart from possible adjustments of a purely technical nature.[6] The published strategy index may incorporate transaction costs or index fees as defined by the index sponsor. Again, the basis for calculation of such fees will remain immutable from the live date of the index onwards.

Investors access alternative beta strategies through a variety of products. Derivatives, in particular total return swaps referencing the value of a strategy index, are the most prevalent investment form, especially for large institutional investors. Investment banks also offer exposure to their strategies via structured notes, funds, and exchange-traded products (ETFs and ETNs). Such investment vehicles will typically carry additional fees over and beyond those incorporated in the strategy index itself.

DATA

Our research uses a sample constructed from a proprietary database of around 3,000 individual strategy indices across seventeen sponsoring investment banks. The database is the property of Clear Alpha Limited, a research and asset management company specializing in alternative beta strategies. The strategies in the database have been recorded since 2010. The aim of the database, since its inception, has been to provide comprehensive mapping of the universe of alternative beta strategies from all the participating investment banks; relationships are maintained with all contributors, and the database is continually updated to reflect new launches and modifications, as well as the cessation of strategies.

The raw database includes numerous duplicate strategies (e.g., the same strategy offered in various currency classes or with different amounts of leverage), and removing the duplicates compresses the data set to around 300 unique strategies. For the purposes of this study, we retained 215 strategies—those for which the index sponsor had indicated the strategy index live date (i.e., the date when the strategy had switched from the backtesting [pro forma] period to the live or "real" returns period). We also removed some very recent strategies that had insufficient data in either the backtest or live period. The minimum length of a backtest period in the final sample is 3.3 years, and the average is 10.7 years. For the live period, the minimum length is 0.44 years and the average is 4.6 years.

Studying this group of 215 indices allows us to identify any changes in risk-adjusted performance prior to and after the live date. The strategies have been developed and sponsored by 15 different investment banks and were commercially promoted between 2005 and 2014. The temporal spread of strategy live dates is reasonably balanced from 2007 onward; there were 96 strategies going live from 2005–2009 and 125 from 2010–2014. All the strategies are denominated in USD and calculated in excess return format (over the short-term risk-free rate) in our analysis. The treatment of the strategies' fees in the database is inconsistent in the cross section (i.e., not all the strategies have fees embedded or charged in a similar manner) but consistent in the individual time series (i.e., the basis of fee charging for a given strategy does not change from backtest to live period).

The original strategy data are daily and begin in January 1990 for the strategy with the longest backtested period. The time series end on March 9, 2015, and all the 215 strategies remain live as of that date. For the purposes of the time series/panel regressions, we truncate any data prior to December 1999 and use monthly returns calculated from the original daily data.

The 215 strategies were classified into five asset classes: commodity (31), equity (69), fixed income (54), FX (40), and multi-asset (21), plus 11 strategy categories based on the strategy description provided

by the sponsoring investment bank. Descriptions of the strategy family definitions used in the remainder of this article are listed in the Appendix.

The market indices, financial instruments, interest, and FX rates used in the analysis have been sourced from Datastream, apart from the JP Morgan Emerging Markets Currency Index, which was sourced from Bloomberg. The primitive trend-following strategy (PTFS) factors of Fung and Hsieh [2001] were sourced from Professor Hsieh's website.[7] The Fama–French global factors, including the global momentum factor, were sourced from Professor French's website.[8]

RESULTS

Exhibit 2 summarizes the key statistics for each strategy group. The results display a drop in excess returns following the live date for all the groups, and most of the groups also show some decline in the standard deviation of returns in the live period. As far as the higher moments of the distribution, we note that all the strategy groups except mean reversion exhibit negative skewness in the live period, as well as excess kurtosis.

Exhibit 3, Panels A–C, illustrates the "realized haircut" (HC_i) in Sharpe ratios, defined as the percentage reduction in SR between the backtested and live periods:

$$HC_i = (SR_{i,Backtest} - SR_{i,Live})/SR_{i,Backest} \quad (1)$$

where $SR_{i,Backtest}$ and $SR_{i,Live}$ are the Sharpe ratios of strategy i during the backtest and live periods, respectively. Some strategies have very lengthy backtest periods; and in light of the findings of McLean and Pontiff [2016], it is possible that some part of the backtest overlaps with the in-sample or pre-academic publication period for the underlying strategy—hence accentuating the reported "historical" returns. Consequently, we also show the SR haircuts using one- and three-year windows immediately prior and after the live date of the strategy. The haircuts shown in Exhibit 3 are summarized across the underlying asset classes, strategy groups, and vintages, respectively, and represent the median haircut of strategies within such asset class, strategy group, or vintage.

The results show a substantial reduction in Sharpe ratios after a strategy goes live; referring back to Exhibit 2, we see that most of the reduction in SR is due to falling excess returns rather than increased risk (volatility). The size of the SR haircut varies somewhat across asset classes and strategy groups, but stands at 50% or more for all the asset classes, and at 60% or more for all strategy groups except volatility when the full-length periods are used. We note that based on our data, the "industry practice" of haircutting backtested SRs by 50% (see Harvey and Liu [2015]) is, if anything, not conservative enough. Exhibit 3, Panel C, illustrates a vintage effect in the data, as strategies with live dates prior to the financial crisis suffer the largest haircuts.

EXHIBIT 2
Descriptive Statistics by Strategy Group

		All	Asset Allocation	Carry	Curve	Event-Driven	Liquidity	Macro	Mean Reversion	Other	Trend Following	Value	Volatility
Average ER p.a.	Backtest	7.80%	7.69%	8.63%	4.78%	9.81%	10.29%	6.53%	8.95%	10.75%	7.22%	9.32%	6.67%
	Live	1.82%	3.65%	0.53%	0.82%	–0.24%	2.98%	0.80%	1.89%	–0.09%	2.38%	0.70%	4.95%
Median ER p.a.	Backtest	6.76%	6.37%	7.72%	3.19%	7.62%	7.75%	6.35%	7.31%	11.60%	5.93%	7.88%	7.57%
	Live	0.97%	4.03%	0.58%	0.52%	–0.51%	1.01%	0.75%	1.11%	–0.38%	0.98%	0.32%	1.55%
Average SD p.a.	Backtest	7.15%	6.60%	6.97%	2.20%	8.63%	9.72%	5.67%	10.19%	7.42%	7.55%	7.60%	7.03%
	Live	6.16%	8.44%	6.03%	1.74%	6.30%	8.08%	6.04%	7.80%	5.73%	6.85%	6.23%	5.84%
Average Skewness	Backtest	–0.22	–0.20	–0.70	–0.10	0.70	0.70	–0.16	1.19	0.06	–0.03	0.03	–2.48
	Live	–0.50	–0.41	–0.99	–0.28	–0.27	–0.23	–0.00	1.02	–0.12	–0.11	–0.47	–3.06
Average Kurtosis	Backtest	18.28	3.45	16.18	9.14	28.78	26.45	5.41	53.70	5.90	6.07	5.43	42.56
	Live	19.62	5.91	22.26	28.67	10.36	32.04	6.10	41.95	3.13	7.16	10.55	37.04
Median Sharpe p.a.	Backtest	1.20	1.21	1.29	1.84	1.09	1.29	1.21	0.89	1.47	1.08	1.12	1.11
	Live	0.31	0.48	0.11	0.31	0.00	0.45	0.20	0.24	0.09	0.38	0.11	0.84
n		215	7	32	21	8	8	13	23	5	53	22	23

Only 18 of the 215 strategies (or 8.4% of the sample) had a live SR equal to or greater than the backtested SR when using the whole sample. In contrast, 65 strategies, or 30% of the sample, had a *negative* SR in the live period. When using the shorter three- and one-year windows surrounding the live date, the results improve somewhat, with 21% and 35% of strategies, respectively, showing an equal or greater SR in the live period than in the backtest period. However, the proportion of strategies with negative SRs doesn't change substantially; it remains at 30% in this inspection.[9]

EXHIBIT 3

Median Sharpe Ratio Haircut by Asset Class, Strategy Group, and Vintage

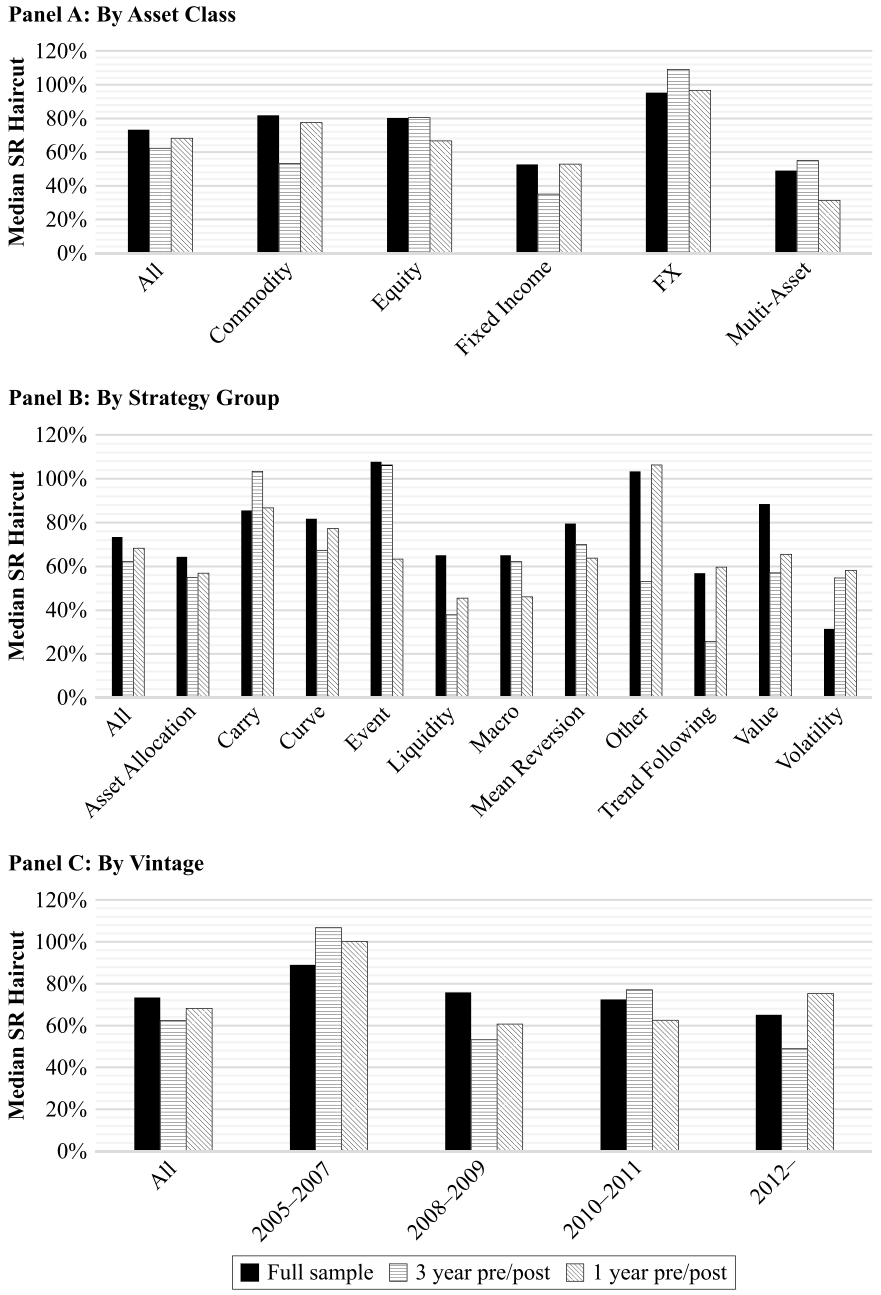


EXHIBIT 4
Cross-Sectional Regressions of Sharpe Ratio Haircut

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Complexity = 2	0.226*	0.149	0.315**	0.229	0.215	0.320**	0.147	0.247*
	(0.137)	(0.132)	(0.131)	(0.141)	(0.139)	(0.137)	(0.133)	(0.143)
Complexity = 3	0.335**	0.347*	0.371***	0.421**	0.318**	0.367**	0.342*	0.429**
	(0.152)	(0.177)	(0.137)	(0.193)	(0.155)	(0.142)	(0.176)	(0.195)
Constant	0.446***	0.373*	0.682***	0.807**	0.613***	0.863***	0.481**	0.964***
	(0.122)	(0.215)	(0.115)	(0.371)	(0.134)	(0.163)	(0.223)	(0.358)
Observations	147	147	147	147	147	147	147	147
R-squared	0.041	0.146	0.169	0.253	0.056	0.178	0.155	0.258
Vintage FE	NO	NO	NO	NO	YES	YES	YES	YES
Strategy FE	NO	YES	NO	YES	NO	NO	YES	YES
Asset Class FE	NO	NO	YES	YES	NO	YES	NO	YES

Notes: Regressions of Sharpe ratio haircut on Complexity dummies and control variables for vintage, strategy group, and asset class.
Robust standard errors in parentheses.
****p<0.01, **p<0.05, *p<0.1.*

We next test a hypothesis following Bailey et al. [2014] that increasing the complexity of a strategy may increase the risk of backtest overfitting, because having a larger number of parameters makes it easier to fit an algorithm to specific historical data. To assess the impact of complexity, we develop a primitive "complexity score" based on the detailed descriptions available on a subset of the strategies in our database. The complexity score takes values from 1 to 3 based on the following criteria:

- Score = 1: Simple strategies with a maximum of one signal. The strategies systematically implement one trade in the same direction, possibly using a wide range of available constituents, without using any endogenous or exogenous signals to either de-risk the trade or, for example, switch the direction of the trade from long to short exposure. An example would be a simple currency carry or short variance swap strategy.
- Score = 2: Medium complexity strategies with one base signal, plus one additional signal such as a risk management overlay or trigger. Any strategies that implement dynamic position sizes or can deleverage their normal position under certain circumstances are included in this group.
- Score = 3: Complex strategies with three or more signals determining the choice of underlying instruments, position direction (long/short), position sizing, or de-risking.

We are able to manually categorize 151 of the 215 strategies according to these criteria, recognizing that there is some qualitative judgment involved in that process. We exclude four of the strategies as outliers; they represent two within the 1st and two within the 99th percentile of the SR haircut distribution. The remaining subset of 147 strategies is reasonably representative of the overall database, with a median SR haircut of 70%, marginally less than the whole sample (73%).

Exhibit 4 reports the results of cross-sectional regressions in which the strategies' realized SR haircuts are explained by the complexity score (dummies for complexity scores 2 and 3). We also include strategy group, asset class, and vintage fixed effects as control variables. The complexity score 3 dummy is statistically significant at the 5% level in most combinations of control variables, and the complexity score 2 dummy is significant in some specifications. The results thus support the proposition that more complex trading strategies are more liable to overfitting and, consequently, deteriorated performance in the live period. After controlling for fixed effects, the most complex strategies appear to suffer from realized SR haircuts that are over 30 percentage points higher than those of the simplest strategies.

We now turn to the common risk factors explaining the performance of the various strategy families. As discussed previously, one of the salient features in the promotion of alternative beta strategies is the ability to get exposure to premia arising from risk factors beyond broad market betas and to extract returns from possible market anomalies or statistical biases. Consequently, as an intuitive and informative experiment, we examine the extent to which alternative beta strategies capture the factor exposures they seek to exploit and whether the exposures remain consistent between backtest and live periods. As a side result, we are able to further analyze the existence of any outperformance (alpha) over and beyond benchmark indices. We report detailed results of pooled panel regressions using the backtest and live periods, respectively, for the equity value strategy and describe the key findings for the three other strategies; in the interest of space, we do not report the regressions statistics.

Equity Value Strategy

We look at the sensitivity of 11 equity value strategies in the database to the three Fama–French [1992] factors (market, value, and size), as well as the momentum factor (Carhart [1997]). According to the strategy sponsor, six of the strategies are "global," two focus on the United States, two on Asia, and one on Europe. Consequently, we use the global version of the four-factor model in the analysis. Exhibit 5 reports our findings.

The results from the backtest period are broadly consistent across the different model specifications. The equity value strategies show positive and statistically significant sensitivity to the Fama–French size and value factors. The constant term in the regressions (alpha) ranges from 47 to 57 basis points (bps) a month and is statistically significant. The live period displays quite a different profile: Somewhat problematically, the value factor becomes statistically insignificant and has a negative sign. The size factor remains positive and statistically significant at the 1% level. There is also a significant positive loading on the momentum factor and market return, and the alpha term becomes statistically insignificant.

Our results are intriguing and somewhat cautionary, as they seem to indicate a failure of the equity value strategy to deliver returns consistent with its stated target and one of the fundamental sources of premia in asset pricing.[10] Furthermore, the strategy appears to generate returns that are more consistent with a long market exposure and a clear momentum bias. This last observation is particularly interesting given the results of Asness et al. [2013]; the authors report a general *negative* correlation between the value and momentum factors in their sample spanning international markets and different asset classes. We also note that our results are consistent with the findings of Glushkov [2015], who reports that over half of equity smart beta ETFs exhibit significant loadings on the size factor.

EXHIBIT 5
Equity Value Strategy

	(1) Backtest	(2) Live	(3) Backtest	(4) Live
Market	0.0104	0.0707***	0.0129	0.0691***
	(0.0212)	(0.0226)	(0.0212)	(0.0232)
Size	0.1150**	0.2460***	0.1130**	0.2440***
	(0.0483)	(0.0578)	(0.0484)	(0.0583)
Value	0.3280***	–0.0207	0.3210***	–0.0182
	(0.0342)	(0.0782)	(0.0340)	(0.0806)
Momentum	0.0174	0.0786***	0.0177	0.0789***
	(0.0187)	(0.0231)	(0.0188)	(0.0234)
Constant	0.0057***	–0.0004	0.0047**	0.0029
	(0.0008)	(0.0010)	(0.0020)	(0.0020)
Observations	1,214	685	1,214	685
R-squared	0.109	0.054	0.122	0.058
Strategy FE	NO	NO	YES	YES

Notes: Regressions of monthly performance of equity value strategies during backtest and live periods, December 1999–February 2015. Market, Size, Value, and Momentum are the Fama–French global market, size and value factors and the momentum factor, respectively.

Robust standard errors in parentheses.

****p<0.01, **p<0.05.*

Equity Volatility Strategy

Our next case study investigates the equity volatility strategy family, which consists of twelve individual products. All the strategies are structurally short volatility—that is, they seek to extract the volatility premium from the market using either option strategies (short straddles or strangles) or short variance swaps. The strategies will typically be exposed to both the changes in implied volatility (short vega) and the realized volatility of the underlying (short gamma).

We start with the Fama–French–Carhart global factors as the explanatory variables for the strategy returns and add two additional factors representing changes in implied equity volatility (the VIX index) and in realized volatility (rolling 20-day realized volatility of the S&P 500 Total Return index). Unsurprisingly, the results show a significant and consistent negative coefficient for both the implied and realized volatility factor both in the backtest and live periods. There is also a consistent, statistically significant positive loading on the momentum factor. Alpha is positive and statistically significant in both the backtest and live periods (39 to 46 basis points monthly outperformance in the live period).

There is one practical caveat to this analysis: Implied volatility, as measured by the VIX index, is not directly investible. As is well known among market practitioners, investible volatility instruments (options or VIX index futures) commonly exhibit a strong contango (rising term structure of volatility). The implication of the contango is that a real-life short volatility strategy would be expected to benefit not only from a positive implied–realized volatility spread, but also from the roll-down effect on a rising volatility futures curve.

To assess whether the positive alpha is simply the result of the roll-down gains not captured by the non-investible VIX index, we include an alternative measure of implied volatility—the iPath S&P 500 VIX Short-Term Futures ETN (VXX:US)—in the regression, which is run from the listing of VXX from February 2009 onward. The results support our suspicion that part of the outperformance of the short-volatility strategies in the baseline regressions is in fact due to the roll-down effect; the alpha term remains positive but reduces from 44 bps a month to 17 bps when VXX returns are used instead of the VIX index change. The alpha also loses its statistical significance once VXX is used in the regression. We conclude that the equity volatility strategy family appears to provide access to volatility risk premia in a reasonably robust manner, but it is not obvious that the strategies outperform a simple exchange-traded short-volatility exposure, on average.

Fixed Income Curve Strategy

Fixed income curve strategies aim to extract returns from long and short positions at different points on a yield curve. We investigate the exposures of the fixed income curve strategy family to the Fung–Hsieh [2001, 2004] hedge fund factors, as well as a simple short-end yield curve slope factor measuring the monthly change in the spread between the USD 2-year interest rate swap and 3-month LIBOR rates. This last factor is intended to serve as a rudimentary measure of (short maturity) term premium.

We find that in the backtest period, the strategy family is negatively loaded to the equity market return (5% level of significance) and the change in the 10-year CMT yield (5% level). The interest rate slope factor is negative but not significant. The backtest alpha is positive and significant (1% level) at 17 to 38 basis points per month, depending on whether strategy-specific fixed effects are included. In the live period, the bond-yield factor remains significant but only at the 10% level, and the curve slope factor coefficient is negative and becomes significant at the 1% level. In addition, the Fung–Hsieh interest rate trend factor is positive and significant. Notably, the alpha remains positive and significant in the regressions for the live period, although its size is roughly halved compared to the backtest period.

The results suggest that the strategy family as a group implemented a long duration trade and benefited from yield curve flattening in the short end. There is some evidence of an interest rate trend-following behavior during the live period; this should probably be seen in the context of the extraordinary monetary policy regime in place during most of the live part of the sample, because official interest rates in major currencies have trended to, or remained at, virtually zero.

FX Carry Strategy

FX carry is perhaps the simplest of the strategy families we encountered in the database. We once again start with the Fung–Hsieh hedge fund factors and then add a further factor representing the returns to a naïve developed markets currency carry strategy; each month, we select the three lowest-yielding currencies as "funding" currencies and the three highest-yielding currencies as the investment currencies, with equal weights. The simple strategy does not include transaction costs.[11] The strategy returns are calculated in USD terms, consistent with our underlying strategy data. The strategy has an SR of 0.73 for the entire period from December 1999 to February 2015—declining from 0.87 during the

average backtest period of the strategies in our database (December 1999–September 2009) to 0.48 during the average live period (September 2009–February 2015). We also include an emerging markets currency index in the regressions.

During the backtest period, the currency carry model is, unsurprisingly, a significant explanatory factor of strategy returns, as is the emerging markets currency index. Alpha is positive and significant at the 1% level, at around 60 bps per month. Turning to the live period, we find a statistically significant (1% level) negative loading on the Fung–Hsieh bond trend factor. As a sign of reasonable relative robustness of the strategy family, the carry model and EM currency factors remain significant (at the 1% level) in the live period, and their factor loadings are almost similar in size to the backtest. Alpha is positive and significant at the 10% level in the model without fixed effects (at 55 bps a month), but becomes insignificant (15 bps per month) when fixed effects are included.

The performance of the simple FX carry model deteriorates during the average live period (September 2009–February 2015), compared to the earlier years. Consequently, we should assign some of the poor absolute results of the strategies in our sample to the lack of performance of the generic strategy—accentuated by the surprise removal of the Swiss franc floor against EUR in January 2015 that caused a −5.4% move in the simple model for the month. Nevertheless, the decline in the average return of the strategies is almost twice as great as that of the naïve benchmark.

DISCUSSION

We recognize that the time period covered in our sample coincides not only with rapid growth in the alternative beta market but also with the extraordinary environment around the global financial crisis of 2008–2009. In an effort to isolate the impact of the crisis, we looked at a subset of 53 strategies that went live prior to 2009, using live data from 2010 onward only. The resulting median SR haircut (relative to the respective backtests) was 81%, or slightly worse than that reported on our full sample (73%). Consequently, we conclude that the financial crisis does not explain the reduction in live performance for the earlier vintages. If anything, it would seem that there is a slight improvement in relative strategy performance (live versus backtest) for the more recent vintages. We discuss the possible explanations for this phenomenon momentarily.

Our results show generally worse performance decay (73% median haircut) than what has been reported in previous research on factor premia (e.g., 26%–58% decline in average returns in McLean and Pontiff [2016]). Several explanations for this finding could exist. It is possible that not all the strategies in our database have been the subject of as rigorous testing as the academically published strategies investigated in previous literature; in other words, some of the strategies could be the result of pure data mining and lack underlying economic rationale. Furthermore, the salience of backtested "historical" returns in strategy marketing further raises the risk of overfitting in the design of the actual investment algorithm.

We highlight again our result regarding the impact of complexity on performance decay, and note that the average Sharpe ratio haircut for the simplest (Complexity score 1) strategies is 44.6%, which is within the range of previous academic findings. Detrimental complexity thus appears to be a plausible reason for the worse performance. We could hypothesize that in addition to unconscious biases arising from lack of portfolio and investment management experience, strategies created by investment banks may have suffered from a focus on "transactional" asset gathering (leading to use of backtested performance as a key marketing tool) at the expense of longer-term maintenance of AUM (robust realized returns). Purely anecdotally, incentives in banking have migrated toward longer-term targets and rewards in the years following the financial crisis, which coincides with a small but noticeable improvement in performance in the more recent strategy vintages.

Discussions with industry participants, including the results in Rabier and Suhonen [2014], indicate a shift in the overall alternative beta market from one dominated by retail and private wealth investors, often accessed via structured products, toward one catering increasingly to sophisticated institutional investors. Such strategic shift could be expected to have a number of effects on the product range. First, institutional fees would most likely be lower than those charged in retail-oriented strategies, causing an improvement in performance of the more recent vintages. Second, institutional investors would be likely to conduct more due diligence on the strategies and place more emphasis on the robustness of sources of premia (risk factors), and to specifically require academic

research and evidence supporting the inclusion of a factor (see e.g., the discussion on desirable properties of factors in Ang et al. [2009] and the equity alternative beta investor survey of Amenc et al. [2015b]). Third, institutions would likely be less reliant on backtested returns as a salient feature when selecting investments. Fourth, industry sources report a trend toward more simplicity in alternative beta strategies—and they increasingly see the role of the strategy developers as providing efficient access to sources of premia rather than devising filters or rules that are liable to introduce biases in live performance (as evidenced by our data). Finally, some retail-oriented alternative beta strategies contain risk control measures, such as investment rules capping the volatility of a strategy. These measures allow attractive pricing of nonlinear payoffs, such as principal protection, but they may be detrimental to the strategy's performance.

The differences in the relative performance of strategies across asset classes raise some interesting questions about the role of banks in the development, execution, and distribution of investment products. With reference to Exhibit 3, Panels A and B, we highlight the relatively poor performance of the FX and equity asset classes, both as far as SR haircuts and the quality of risk-adjusted returns are concerned. Furthermore, we note the weak results regarding the value strategy and the lack of factor robustness in the equity value family. Several possible explanations exist for the poor performance of equity value and the equity asset class overall. The strategy family may suffer from a general lack of robustness and poor out-of-sample performance, as proposed by Glushkov [2015], Amenc et al. [2015a], and McLean and Pontiff [2016], among others. It is also possible that the results reflect different ways of defining and measuring "value" in equities among market practitioners.

The plethora of investment funds active in the equity markets ranging from passive index funds and "smart beta" ETFs to mutual funds and hedge funds adds competitive pressure not only to return generation but also to the marketing of strategies, which may result in data mining and unnecessary complications of algorithms in order to improve backtested performance. The prominent presence of retail investors in the equity markets relative to other asset classes, and especially through structured products,[12] is noteworthy in this context.

On the other hand, our results indicate more robust performance in terms of live versus backtest periods and significance of SRs and stability of factor exposures in the fixed income asset class and the volatility and curve strategy groups. Our suggested interpretation of the data is that these strategies, and fixed income as an asset class, are structurally more institutionally focused, and consequently they would benefit from the market trends discussed. In particular, we would hypothesize that from an investor's perspective, there is value in getting efficient exposure to investment strategies such as the fixed income curve and equity volatility that have solid theoretical and empirical underpinnings, but which not straightforward to implement without a reasonable level of market knowledge, systems and trading, settlement, collateral, and risk management infrastructure. Hence, the need for attempted value-adding from investment banks in the form of complex investment algorithms and filters in these strategies could be less pronounced.

We would also note that because alternative beta development and trading is often part of a structured derivatives business in an investment bank, there is likely to be a strong quantitative bias to the sponsor's skillset and strategy construction process, possibly at the expense of a more qualitative assessment of the strategies and their theoretical foundations. Such focus on quantitative methods could be a more appropriate approach in fixed income and volatility investing than in other strategies.

CONCLUSIONS

This article has presented empirical evidence of the performance and robustness of alternative beta strategies across asset classes and strategy groups. Our results support the recent warnings in finance literature regarding "factor fishing," multiple testing, overfitting, and selection and reporting biases in financial research and product development. The findings highlight the importance of detailed due diligence on quantitative strategies and suggest that backtested performance and risk measures may offer limited value to practical alternative beta strategy selection and portfolio management.

On a positive note, we find some evidence of improved strategy robustness in the later years of our sample, possibly reflecting the requirements of a more institutional client base. Furthermore, asset classes other than equities appear to be less adversely impacted by backtest biases. Consequently, we see a possible future agenda for the financial services industry in the offering

of efficient access to robust, academically and empirically justifiable factor exposures across asset classes (and in particular beyond equities) using minimal filtering or other elaborate investment rules.

Our results show that business practices surrounding alternative beta strategies can be improved, and the strategies themselves made more reliable—both in terms of development and disclosure of key information to potential investors. We point to the fact that this study regarding the backtest bias in alternative beta strategies falls within a broader scope of business improvements that have been highlighted by regulators in recent years. More specifically, we encourage readers to consult the International Organization of Securities Commissions' Principles for Financial Benchmarks [2013] (under the definition of which alternative beta strategies fall) and the U.K. Financial Conduct Authority's recent thematic review on financial benchmark activities [2015].

APPENDIX

STRATEGY DEFINITIONS

We have classified the 215 individual strategies into 11 strategy groups across five asset classes. The definitions for the strategy groups are as follows. The numbers in parentheses show the number of individual strategies within each group.

Asset Allocation (7): Strategies implementing allocation across several asset classes according to systematic rules and either endogenous or exogenous rebalancing signals.

Carry (32): Strategies seeking excess returns by going long higher-yielding assets and simultaneously going short lower-yielding assets.

Curve (21): Strategies generally aiming to earn a term premium by investing in longer maturities or taking other systematic positions along a yield or futures curve.

Event-Driven (8): Strategies seeking to exploit asset price moves or convergence following specific events such as mergers and acquisitions or earnings announcements.

Liquidity (8): Strategies seeking to capture a liquidity premium in asset prices either via long/short positions in less versus more liquid assets of otherwise comparable risk characteristics or via liquidity provisions to the market.

Macro (13): Strategies allocating to investments within an asset class using external macroeconomic indicators.

Mean Reversion (23): Strategies seeking to exploit mean reversion in asset prices by selling recent outperformers and buying underperformers or by using non-delta hedged option-selling strategies.

Other (5): Primarily multistrategy products that could not be classified to the other categories.

Trend Following (53): Strategies taking positions in target assets according to recent price movements, typically seeking to identify and follow trending markets.

Value (22): Strategies seeking to invest in relatively undervalued assets and short (or underweight) overvalued assets, based on fundamental valuation measures.

Volatility (23): Strategies seeking to extract the premium between implied and realized asset volatility.

ENDNOTES

We would like to thank an anonymous reviewer as well as Tim Edwards, Antti Ilmanen, Matti Keloharju, Chi Lee, Mikko Niemenmaa, and Vesa Puttonen for their helpful comments and suggestions. Any errors are our own.

[1]Terminology in the market is far from standardized, but "smart beta" typically refers to long-only investment products that use alternative weighting methods to market-capitalization weighted indices. Various titles, including alternative, quantitative, or efficient beta, or factor, style, or alternative risk premia are used to define the broader investment universe incorporating long–short strategies and leverage. This article uses the term *alternative beta* to encompass quantitative, nondiscretionary investment strategies that may include short exposures and leverage.

[2]For example, Rabier and Suhonen [2014].

[3]For practical debate, see, e.g., Dickson et al. [2012], Beddall and Land [2013], and Evans and Schmitz [2015].

[4]For a practitioner-oriented discussion on the persistence of known factors returns, see e.g., Asness [2015] and Preston et al. [2015].

[5]Overview based on survey in Rabier and Suhonen [2014].

[6]It is possible that a new strategy is initially only marketed to a small group of investors during an "incubation" period after the live date; and only if the performance remains positive thereafter, the strategy is selected for broader, public marketing. To the extent that such selection occurs, the universe of publicly launched strategies could be argued to be subject to survivorship bias. Consequently, our results may be somewhat too optimistic so far as realized performance is concerned. See, for example, Evans [2010] for analysis of incubation in mutual funds.

[7]http://faculty.fuqua.duke.edu/~dah7/DataLibrary/TF-FAC.xls.

[8]http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[9]We also examined the statistical significance of the SR in the backtest and live periods (see Sharpe [1994]). Ninety-

five percent of the Sharpe ratios of the 215 strategies in our sample are positive and statistically significant at the 10% level (i.e., p-value < 0.1), and 80% are significant at the 1% level in the backtest period. In the live period, only a quarter of the strategies (26.5% to be exact) have an SR different from zero at the 10% level of statistical significance, and only 6% or 13 strategies are significant at the 1% level. In addition to the vintage effect—poor performance of the pre-crisis vintages and above-average performance for the class of 2008–2009—we find a particularly poor live risk-adjusted performance in the FX and equity asset classes and event-driven, value, mean reversion, and carry strategy groups. On the positive side, almost half the strategies in the fixed income asset class have live SRs that are positive and significantly different from zero.

[10]Because of multicollinearity between the factors, the other factors could "outperform" value, even if value had significance on its own. We test this possibility by removing the size and momentum factors from the regressions, but the results remain essentially unchanged.

[11]The underlying currencies are AUD, CAD, CHF, EUR, GBP, JPY, NOK, NZD, SEK, and USD. The model was inspired by Ilmanen [2011], who notes that the inclusion of transaction costs representative of the market environment in the 2000s is unlikely to detract more than 0.1 from the SR of the strategy.

[12]Célérier and Vallée [2014] report that 70.1% of the 55,000 structured products launched in 17 European countries in 2002–2010 were linked to equity underlyings.

REFERENCES

Amenc, N., F. Goltz, A. Lodh, and S. Sivasubramanian. "Robustness of Smart Beta Strategies." An ERI Scientific Beta Publication, 2015a.

Amenc, N., F. Goltz, V. Le Sourd, and A. Lodh. "Alternative Equity Beta Investing: A Survey." An EDHEC-Risk Institute Publication, 2015b.

Ang, A. *Asset Management. A Systematic Approach to Factor Investing.* Oxford University Press, 2014.

Ang, A., W.N. Goetzmann, and S.M. Schaefer. "Evaluation of Active Management at Norwegian Government Pension Fund—Global," 2009.

Asness, C.S. "How Can a Strategy Still Work If Everyone Knows about It?" *Institutional Investor*, September 2015.

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

Bailey, D.H., and M. Lopez de Prado. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *The Journal of Portfolio Management*, 40 (2014), pp. 94-107.

Bailey, D.H., J.M. Borwein, M. Lopez de Prado, and Q.K. Zhu. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Fitting on Out-of-Sample Performance." *Notices of the AMS*, 61 (2014), pp. 458-471.

Beddall, M., and K. Land. "Hypothetical Performance of CTAs." Winton Capital Management, 2013.

Bioy, H., A. Bryan, J. Choy, J. Gabriel, B. Johnson, S. Lee, T. Murphy, A. Prineas, and G. Rose. "A Global Guide to Strategic-Beta Exchange-Traded Products." Morningstar, 2015.

Carhart, M.M. "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52 (1997), pp. 57-82.

Célérier, C., and B. Vallée. "The Motives for Financial Complexity: An Empirical Investigation." Working paper, 2014.

Dickson, J.M., S. Padmawar, and S. Hammer. "Joined at the Hip: ETF and Index Development." Vanguard Research, 2012.

Evans, R.B. "Mutual Fund Incubation." *Journal of Finance*, 65 (2010), pp. 1581-1611.

Evans, A., and C. Schmitz. "Value, Size and Momentum in Equity Indices—A Likely Example of Selection Bias." Winton Capital Management, 2015.

Fama, E.F., and K.R. French. "The Cross-Section of Expected Stock Returns." *Journal of Finance*, 47 (1992), pp. 427-465.

Financial Conduct Authority. "Financial Benchmarks: Thematic Review of Oversights and Control." U.K. Financial Conduct Authority," 2015.

Fung, W., and D.A. Hsieh. "The Risk In Hedge Fund Strategies: Theory and Evidence from Trend Followers." *Review of Financial Studies*, 14 (2001), pp. 313-341.

———. "Hedge Fund Benchmarks: A Risk-Based Approach." *Financial Analysts Journal*, 60 (2004), pp. 65-80.

Glushkov, D. "How Smart Are Smart Beta ETFs? Analysis of Relative Performance and Factor Exposure." Working paper, 2015.

Harvey, C.R., and Y. Liu. "Backtesting." *The Journal of Portfolio Management*, 42 (2015), pp. 13-28.

——. "Evaluating Trading Strategies." *The Journal of Portfolio Management*, Vol. 40, No. 5 (2014), pp. 108-118.

Harvey, C.R., Y. Liu, and H. Zhu. "… And the Cross-Section of Expected Returns." Working paper, 2014.

Ilmanen, A. *Expected Returns: An Investor's Guide to Harvesting Market Rewards*. West Sussex, U.K.: John Wiley and Sons, Ltd, 2011.

International Organization of Securities Commissions. "Principles for Financial Benchmarks Final Report." The Board of the International Organization of Securities Commissions, 2013.

McLean, R.D., and J. Pontiff. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, 71 (2016), pp. 5-32.

Preston, H., T. Edwards, and C.J. Lazzara. "The Persistence of Smart Beta." S&P Dow Jones Indices Research, 2015.

Rabier, B., and A. Suhonen. "Quantitative Investment Survey: Beyond Fundamental Indexing." Allenbridge Investment Solutions, LLP, 2014.

Sharpe, W.F. "The Sharpe Ratio." *The Journal of Portfolio Management*, 21 (1994), pp. 49-58.