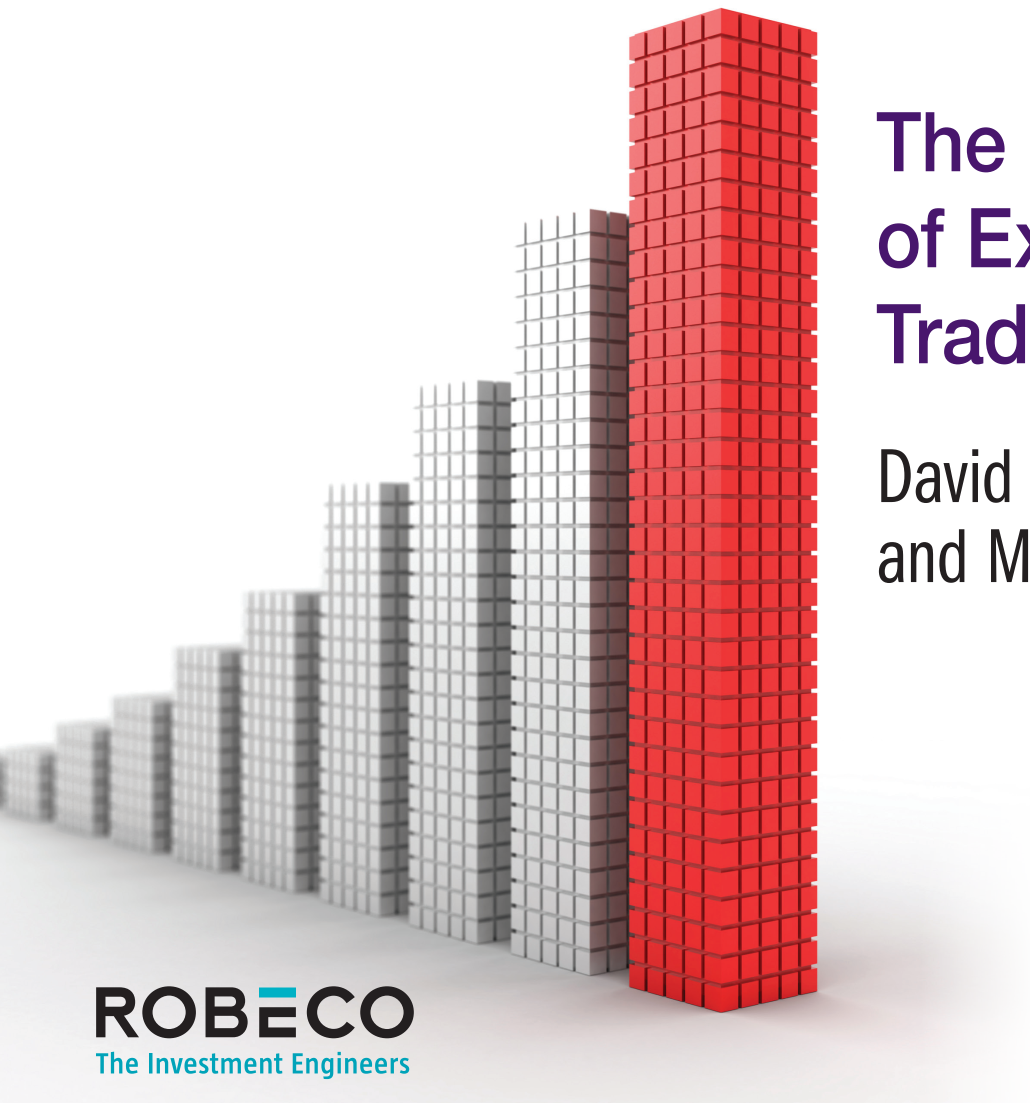
PORTFOLIO MANAGEMENT RESEARCH

The Journal of Alternative Investments

Winter 2021 – Volume 23 Issue 3 | JAI.pm-research.com



The Performance of Exchange-Traded Funds

David Blitz
and Milan Vidojevic

Official Publication of the CAIA Association®

The Performance of Exchange-Traded Funds

David Blitz and Milan Vidojevic

David Blitz
is the head of quantitative research at Robeco in Rotterdam, the Netherlands.
d.c.blitz@robeco.nl

Milan Vidojevic
is a senior quantitative researcher at Robeco in New York, NY and is affiliated with Vrije Universiteit in Amsterdam, the Netherlands.
m.vidojevic@robeco.com

KEY FINDINGS

- ETFs have collectively lagged the market by about the same amount as active mutual funds.
- Most smart beta ETFs have also failed to beat the market.
- From a pure performance perspective, the allure of ETFs finds little support in the data.

ABSTRACT

Exchange-traded funds (ETFs) are commonly regarded as an efficient, low-cost alternative to actively managed mutual funds, yet their perceived superiority is largely anecdotal. This article evaluates the performance of a comprehensive, survivorship-bias-free sample of US equity ETFs following the approach that has been commonly used to evaluate the performance of actively managed mutual funds. The authors find that ETFs have collectively lagged the market by an amount similar to the widely documented underperformance of active mutual funds. They perform textual and regression-based analysis to identify factor ETFs and show that most of these have also failed to beat the market. They conclude that from a pure performance perspective, the allure of ETFs finds little support in the data.

TOPICS

*Factor-based models, mutual fund performance, passive strategies, exchange-traded funds and applications**

*All articles are now categorized by topics and subtopics. View at PM-Research.com.

The performance of actively managed mutual funds has been the subject of intense scrutiny in the academic literature. The seminal Carhart (1997) study finds that actively managed mutual funds on average underperform the market after management fees and transaction costs and that there is no evidence for persistence in performance after controlling for exposures to the market, size, value, and momentum factors. Subsequent studies have generally confirmed these findings (see, for instance, Fama and French 2010), resulting in a broad consensus in the academic community that investors should reallocate drastically from actively managed funds to passive replication of the broad market portfolio (see French 2008). The investment community has also embraced these concepts, with even the world-renowned stock picker Warren Buffett advocating the use of S&P 500 Index funds for the “average” investor.

Passive investing has been facilitated by and popularized following the introduction of exchange-traded funds, or ETFs for short. These funds are continuously traded on stock exchanges and typically aim to passively replicate a well-known index at low costs. Over the past two decades, ETFs have grown spectacularly, which raises the question whether conventional mutual funds are being rendered obsolete. Without supporting evidence, however, it is premature to jump to this conclusion. One reason for caution is that the main differentiator of ETFs, continuous trading, should be of little relevance to passive investors because the whole idea of the passive approach is to buy and hold for the long term and refrain from trading altogether. A second consideration is that not every ETF involves low costs. Whereas the cheapest ETFs have annual expense ratios below 0.05%, there are also ETFs with expense ratios above 1%, which makes them more expensive than many mutual funds. Third, if the purpose of ETFs were to facilitate passive investing, then in theory, one ETF on the broad market portfolio would suffice. In reality, one would expect perhaps a few more funds because of practical matters, such as competition between different providers, asset classes, or time zones, but not thousands of funds. While there is a handful of very big ETFs that track a broad market index, such as the S&P 500, the vast majority of ETFs track indices that themselves represent active strategies.

Whereas the performance of the traditional mutual funds has been extensively investigated in the literature, there is not much known yet about the realized performance of ETFs. This article aims to fill this gap in the literature by conducting a performance analysis of a comprehensive, survivorship-bias-free sample of the US-listed ETFs investing in US equities. Our sample consists of over 900 ETFs, with almost $1.9 trillion under management at the end of the sample period (December 2017). We evaluate the performance of ETFs by applying the same standards that are commonly applied to mutual funds, which is to compare them with the CRSP market portfolio and the standard Fama–French–Carhart (FFC) factors. One might object that these factors ignore fees and costs, but this is how the bar has been set in the existing literature on mutual fund performance. We also consider a modified FFC six-factor model with factors that better reflect investable strategies. We find that the modified FFC model is more suitable for the purpose of performance evaluation, but that the main conclusion drawn with the standard model remains unchanged.

In Exhibit 1, we present the percentages of ETFs with higher and lower returns, volatilities, and Sharpe ratios than the capitalization-weighted CRSP market portfolio over their entire live period, taking all ETFs for which at least 12 consecutive monthly returns are available. Based on realized returns, 60% of ETFs underperformed the market, 80% exhibited higher volatility, and 80% underperformed in terms of Sharpe ratios. Such figures do not appear to be much different from what has been reported for actively managed mutual funds.

Exhibit 2 shows the cumulative CAPM alpha of an asset-weighted portfolio of all ETFs in our sample. The start date for this analysis is the end of 2003, which is the first year-end at which the number of ETFs exceeded 100. The graph shows that, collectively, ETFs have substantially underperformed the overall equity market, with an annualized CAPM alpha amounting to –0.75%. Fama and French (2010) report that the asset-weighted portfolio of all US actively managed mutual funds generates returns before expenses in line with those of the market, and negative returns after fund expenses are taken into account. On a net return basis, mutual funds in their sample generate a CAPM alpha of –1.13% a year. Because fees and expenses have come down over time and our ETF study covers a more recent period, the gap in performance between ETFs and mutual funds does not appear to be very large.

Zooming in on the different types of ETFs, we find that a small number of generally big ETFs, which aim to track one of the broad market indices, live up to their promises. We do observe that by following popular indices, such as the S&P 500 and Russell 1000, rather than the true broad market portfolio, many of these funds only provide exposure to the large-cap segment of the market.

The weak overall performance of ETFs turns out to be mainly driven by the large number of ETFs that do not aim to replicate any of the broad market indices.

EXHIBIT 1

Breakdown of ETF Performance Characteristics versus the Market

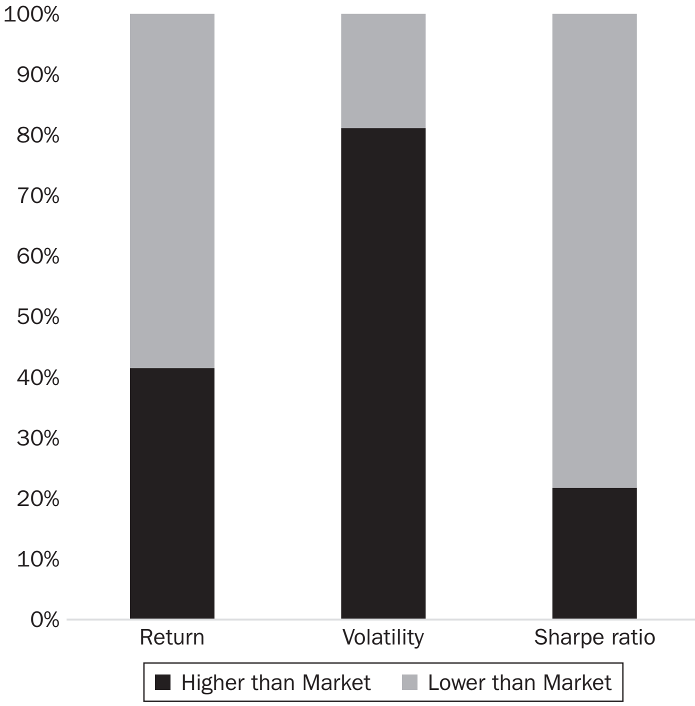


EXHIBIT 2

Cumulative CAPM Alpha of All ETFs

Within this group, we find particularly dismal long-term performance for the leveraged and inverse equity ETFs, which suggests that these funds are merely interesting for short-term speculative purposes, as also argued by Cheng and Madhavan (2009). The poor long-term performance of the leveraged and inverse equity ETFs likely stems from the fact that they require active, daily rebalancing of their derivative positions to maintain their targeted exposures.

Among the remaining ETFs—that is, all ETFs excluding those tracking the broad market indices and leveraged and inverse equity ETFs—we focus on the performance of funds that explicitly or implicitly provide exposures to the factors that are well established in the asset-pricing literature. We refer to these funds as *factor* ETFs and identify them based on their names by performing a keyword search and, alternatively, using time-series regressions on the academic factor return series. The performance of factor ETFs turns out to be disappointing, with Sharpe ratios lagging the market across the board and most CAPM and multi-factor alphas being negative. The magnitude of these alphas again appears to be quite similar to what one might expect from conventional actively managed funds. Our finding of poor performance for factor ETFs is in line with Gluskov (2016), who examines the performance of a smaller sample of smart beta ETFs. It must be noted, however, that the recent environment for some of the prominent factor strategies has not been favorable, as realized performance of the classic academic factors in the United States over the last 10–15 years has significantly lagged their longer-term realizations. Given that some factor ETFs do provide large and significant exposures to the targeted factors, they can be expected to add value if factor premiums rebound in the future. A caveat here is that the factor exposures of some ETFs may have been obtained unintentionally, which means that these exposures might change in the future.

In addition to forming factor ETF portfolios, we also create anti-factor ETF portfolios by selecting ETFs with negative and statistically significant exposures to the academic factor premiums. We find that three out of the six groups have significantly underperformed the market and the other three have returns in line with the market. Therefore, although it has been hard to beat the market by investing in ETFs with good factor exposures, it would have at least been wise to avoid the ones with bad factor exposures in order to prevent lagging the market.

Altogether our examination of US equity ETFs shows that performance is not as impressive as one might expect it to be, as investors in these ETFs have collectively realized a performance that does not appear to be much different from the widely documented performance of conventional, actively managed mutual funds.

ETF BACKGROUND, DATA, AND CLASSIFICATION

ETF Background

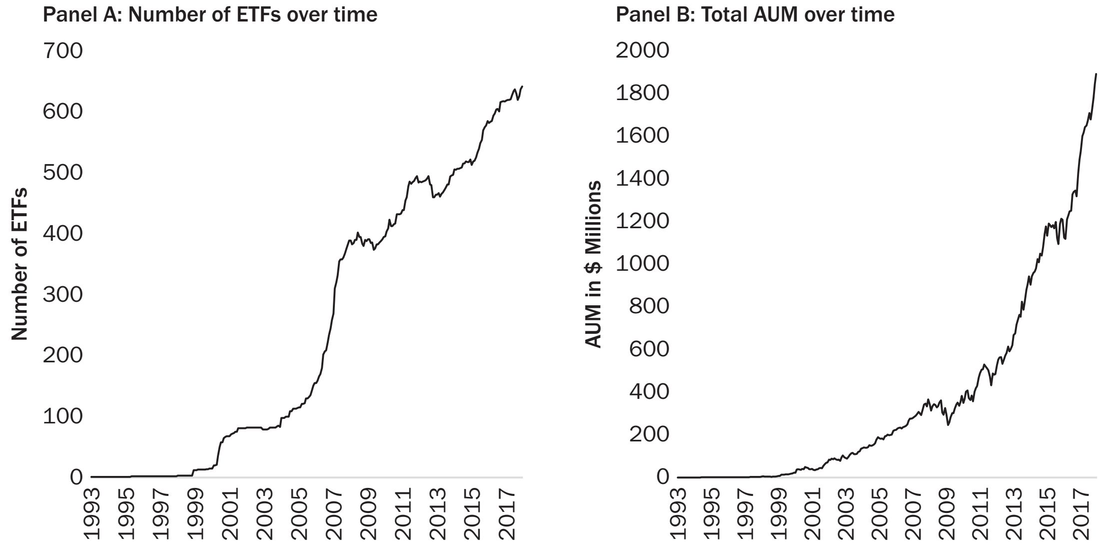
Passive index funds were first introduced by Vanguard in 1975 and their market share has steadily grown since. Nowadays, investors have sizable allocations to passive index strategies, and for many, it has even become the default investment approach. The adoption of passive index strategies has accelerated since the introduction of the first exchange-traded fund (ETF) in 1993, a fund that seeks to replicate returns of the S&P 500 at low costs. At the end of 2017, the total number of ETFs globally was over 4,500, with a combined market value exceeding $4 trillion.[1] ETFs share many similarities with conventional mutual funds but differ in the way they are cleared. Whereas mutual funds can be bought or sold just once a day, ETFs are traded throughout the day on stock exchanges, similar to common stocks. An important caveat here is that the market prices of ETFs can deviate substantially from their NAVs despite the arbitrage mechanisms that are supposed to ensure the market prices stay close to the value of the underlying assets; see Petajisto (2016). An extensive stream of literature has compared ETFs and mutual funds on operational efficiency, costs, liquidity considerations, and the impact on financial markets in general. For an overview of this literature, we refer the reader to Ben-David, Franzoni, and Moussawi (2017).

Traditionally, ETFs have aimed to efficiently capture the equity market return by replicating the performance of a broad capitalization-weighted market index at low fees. For this reason, many consider the terms passive and ETF to be synonyms. This perception is not correct, however, because ETFs themselves are not investment strategies, but vehicles (market instruments) used to implement set strategies. Most ETFs, in fact, do not track broad market indices but indices focusing on a certain sector or investment theme (value, growth, small-cap, income, etc.). These ETFs use passive investment techniques to replicate indices that themselves reflect an active departure from the broad market portfolio, targeting investors who explicitly or implicitly wish to express active investment views. Some ETFs are even outright actively managed. Easley et al. (2018) find that ETFs have a median active share of 93.1% and median tracking error of 8.8%, relative to the passive market portfolio. Robertson (2018) concludes that far from being passive, ETFs are better understood as a different form of delegated management, where the delegee is the index creator rather than the fund manager.

Ben-David, Franzoni, and Moussawi (2018) find that investors in ETFs have substantially shorter investment horizons than investors in mutual funds. This explains why ETFs are sometimes referred to as the new derivatives. For instance, a range of sector ETFs facilitates or even encourages an active sector rotation approach. In support of this concern, Bhattacharya et al. (2017) find that poor timing decisions by private investors in Germany completely negate the benefits of efficient, low-cost security selection by ETFs.

[1] Source: https://www.statista.com/topics/2365/exchange-traded-funds/.

EXHIBIT 3
Sample Description



ETF Data

Our sample consists of all ETFs that invest in US equities that are listed on the US exchanges (NYSE Arca/ArcaEx and NASDAQ—exchange codes 4 and 3, respectively, and share code 73). Monthly total returns, shares outstanding, and closing prices are obtained from the CRSP database. The securities are further matched with the Bloomberg and FactSet databases on their CUSIPs to obtain information on the fund name, asset class, industry focus, market-cap focus (e.g., large or small cap), strategy, and geographic focus. Data availability starts in January 1993, when State Street Global Advisors launched the first ETF, the S&P 500 Trust ETF. This fund remained the largest one throughout our sample period that ends in December 2017. Throughout the article, we construct value-weighted (i.e., AUM-weighted) portfolios of ETFs. In unreported tests, we find that all our results continue to hold if we use equal weights or if we use value weights and cap the weight of any given ETF to 10% and distribute the excess weight proportionally across all other ETFs in the portfolio.

The sample consists of 918 matched ETFs, 276 of which are no longer in existence at the end of the sample period. These "dead" ETFs are included in our analysis to ensure that survivorship bias issues do not distort the results. At the end of the sample, there are 642 live ETFs with combined assets under management (AUM) of just under $1.9 trillion. Further, 181 of these ETFs have existed for less than three years. Exhibit 3 shows how the number of ETFs (Panel A) and their total AUM (Panel B) has developed over time. The strong growth of the ETF market over the last couple of decades is clearly visible in these graphs.

The distribution of ETF AUM is highly skewed. Exhibit 4 shows its 25th, 50th, 75th, and 95th percentiles at the year-end for the last 10 years. We also show the AUM of the largest ETF in the universe (SPDR). At the end of the sample, this single ETF accounts for 14.7% of the AUM of the entire universe, and the largest 50 ETFs account for 80%.

We obtain the factor return series and the risk-free rate from the online database of Professor Kenneth French. We consider the five factors that comprise the Fama and French (2015) five-factor model—market (Mkt-RF), size (SMB), value (HML), profitability (RMW), investment (CMA)—and we augment them with the momentum

EXHIBIT 4

AUM Distribution at the End of the Sample (millions USD)

	Percentile				
	25th	50th	75th	95th	Largest
2007	10	56	330	3,583	100,826
2008	12	61	308	3,176	94,893
2009	24	88	417	3,739	84,865
2010	23	109	418	4,838	90,965
2011	14	75	355	4,871	87,195
2012	20	102	483	6,530	123,695
2013	33	156	716	8,394	174,686
2014	43	206	935	10,273	215,916
2015	27	156	750	10,066	183,999
2016	28	182	866	11,467	224,621
2017	33	188	938	12,702	278,344

EXHIBIT 5

Performance of Academic Factor Portfolios in the United States, 2004–2017

	Mkt-RF	SMB	HML	RMW	CMA	WML
Return	8.54%	1.54%	0.41%	3.57%	–0.23%	0.83%
Volatility	13.78%	8.20%	8.71%	5.65%	4.86%	15.35%
Sharpe ratio	0.62	0.19	0.05	0.63	–0.05	0.05
CAPM alpha		–0.58%	–1.07%	5.18%**	–0.21%	3.93%

NOTES: **Statistically significant at the 5% level, *Statistically significant at the 10% level.

factor (WML). Exhibit 5 shows the performance of the factor premiums over our main research sample period (2004–2017). We see that only one out of the five factors, RMW, generated a significant CAPM alpha, confirming the result of Arnott et al. (2019) that factor premiums have failed to deliver over the last 15 years. If factor premiums fail to materialize, then we expect to see this reflected in the performance of ETFs, which provide exposure to these factors.

ETF Classification

We start by examining the aggregate performance of the ETFs in our sample, similar to how the collective performance of conventional mutual funds has been evaluated in the literature. We acknowledge that ETFs are used by many different investors for many different purposes, but this is not fundamentally different from conventional mutual funds. For instance, many ETFs are designed to provide exposure to a certain industry, theme, or style, but so are many mutual funds. And just as ETFs can be used to implement a tactical view or provide a hedge within a broader portfolio, so too can mutual funds. We also acknowledge that the aggregate performance of ETFs is affected by the timing decisions of the investors in ETFs, but again, this concern also applies to mutual funds. Looking at the combined performance of all ETFs allows us to assess how much the entire investment community has been better or worse off as a result of investing in ETFs.

We next classify funds into groups in a number of ways. First, we broadly group funds into market index trackers, inverse and leveraged equity ETFs, and a group that consists of all remaining ETFs. We define the broad market trackers as all ETFs with full sample annualized tracking errors with respect to any of the five commonly used equity market indices (CRSP total market portfolio, S&P 500, Russell 1000, Russell 3000, and Dow Jones Industrial Average) under 1%. While a tracking error against an index of 1% seems high in today's environment, historically, ETFs have not been as accurate at tracking indices as they are now. A manual check confirms that the selected ETFs are indeed tracking the broad market indices. The leveraged and inverse equity ETFs are categorized based on the fund category provided by Bloomberg and FactSet. Panel A of Exhibit 6 shows the number of ETFs across these groups over time. We see that a very small number of funds are classified as the broad market trackers, only 14 over the entire sample period, and that the number of leveraged and inverse equity ETFs has sharply increased between the years 2006 and 2010, after which it has leveled off.

Within the group of all remaining ETFs, we create subsets of ETFs that either explicitly, based on their name, or implicitly, based on exposures inferred from their returns, provide exposures to well-known factor premiums. A stylized illustration of our classification approach is given in Exhibit 7.

For the name-based classification, we conduct a keyword search in ETF names for the words listed in the following table:

Classification (total number)	**Keyword**
Small (91)	Small
Value (70)	Value, fundamental, RAFI
Dividend (72)	Dividend, income, revenue
Momentum (25)	Momentum
Quality (15)	Quality
Low risk (21)	Low volatility, minimum volatility/ variance, low beta

The screening results are further manually verified. For the ETFs in the small-cap and dividend categories, we also augment the groups with ETFs that do not necessarily have the classifying words in their name but have small-cap or dividend as their strategy focus category on Bloomberg. Panel B of Exhibit 6 shows how the number of ETFs within each group evolves over time. The small-cap, value, and dividend-style ETFs tend to be the most represented in the sample.

For the classification inferred from the realized ETF returns, we perform return-based style analysis for each ETF over its entire history. Specifically, we regress the excess returns of an ETF on the market, size, value, momentum, profitability, and investment factors in order to determine whether it has provided exposure to any of these factors. We require at least 36 consecutive monthly return observations, and an ETF is considered to be a factor fund if the coefficient on the respective factor has a t-statistic of greater than 2. In order to classify the low-risk ETFs, we run regressions of ETF returns in excess of the market return on the market factor and select the ETFs with statistically significant, negative betas. Monthly factor returns are obtained from the online data library of Professor Kenneth French. Panel C of Exhibit 6 shows the number of ETFs in each exposure group. Not surprisingly, more ETFs are classified using the regression approach than using the name-based classification because funds can have significant exposures to factors that are not explicitly targeted. The overlap between the name-based and exposure-based groups based on the same factor is high: 86.7% for value, 96.3% for size, 64.3% for momentum, 77.8% for quality, and 66.7% for low risk. However, a non-trivial fraction of the factor ETFs identified by name fail to provide significant exposures to the targeted factor.

EXHIBIT 6

Number of ETFs

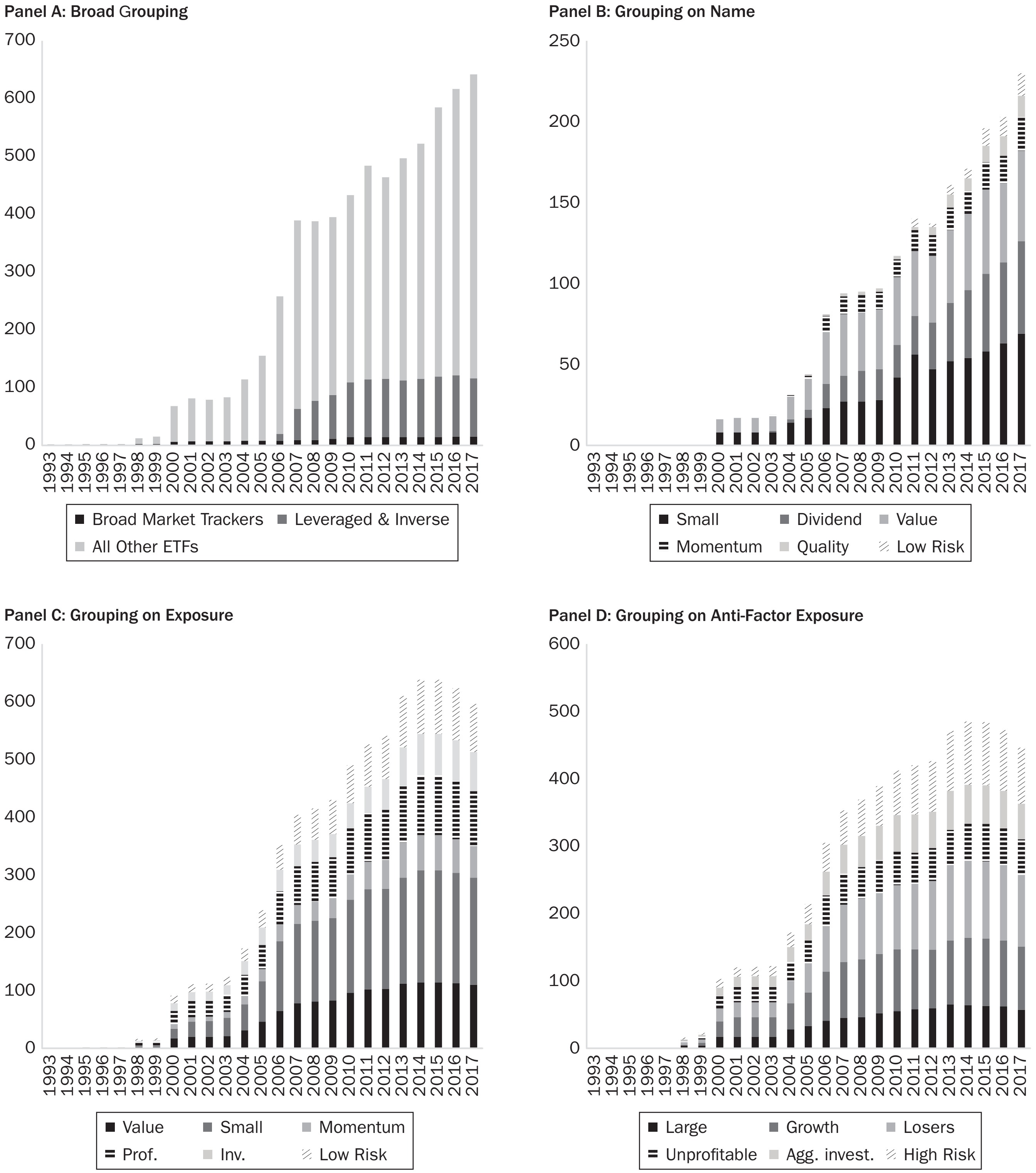
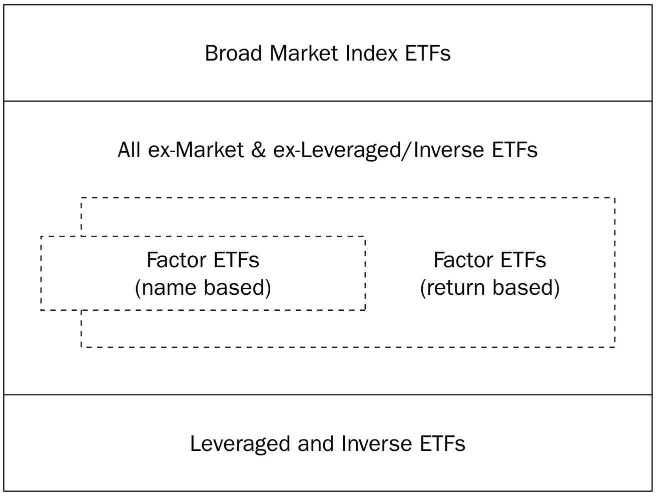


EXHIBIT 7

A Styled Illustration of Our Classification Approach



In addition to the factor ETFs, we form anti-factor ETF groups by selecting funds with negative and statistically significant exposures to the academic factors. This yields six groups: large, growth, losers, unprofitable, aggressive investment, and high risk. For the high-risk group, we apply the same approach that we used to classify the low-risk ETFs, but now we select the funds with positive and significant, as opposed to negative and significant, betas in excess of the market. Panel D shows that the number of ETFs in each category is quite sizable and comparable to what we reported for the factor ETFs. We note that some ETFs in Panels B, C, and D can appear in multiple groups as they contain multiple keywords in their name (e.g., small-cap value appears both in the small and in the value group) or have significant exposures to more than one factor.

In the remainder of this study, we focus on the sample period from January 2004 until December 2017. Before this period, the total number of ETFs was less than 100, and many factor categories either did not exist or consisted of a very small number of funds. We note that due to the short sample period, the statistical significance of our results is oftentimes weak.

THE PERFORMANCE OF ETFs

Performance of Broad ETF Categories

The first column of Exhibit 8 compares the aggregate performance of all ETFs with the performance of the overall equity market portfolio. The equity market is a commonly used benchmark in the studies that evaluate the performance of conventional actively managed mutual funds, and therefore we subject the ETFs to the same level of scrutiny. Over the 2004–2017 sample period, the raw return of ETFs lagged the market by 0.91% per annum. On a risk-adjusted basis, we also observe an underperformance, with a Sharpe ratio of 0.56 versus 0.62 for the market and a statistically significant CAPM alpha of –0.75% per annum.

We proceed by making a distinction between the ETFs that passively track a broad market index and the rest. Although the broad market index ETFs are small in number, they are quite big in terms of their assets under management. The second column of Exhibit 8 shows that the broad market index ETFs underperformed the CRSP market index by 0.47% per annum, but their volatility was also lower. On a risk-adjusted basis, we observe a slightly lower Sharpe ratio of 0.60 versus 0.62 for the market and a CAPM alpha of –0.18% per annum, which is, however, statistically indistinguishable from zero. The main cause of performance deviations between these funds and the market portfolio is a systematic underweight of small-cap stocks, because the most popular indices, such as the S&P 500 and Russell 1000, only consist of large and mid-cap stocks. Although the six-factor alpha can account for such exposures, it amounts to –0.44% per annum, which is statistically significant at the 5% level.

The group of all other ETFs exhibits a much less impressive performance. The aggregated raw return of this group falls short of the market by more than 1% per annum, the Sharpe ratio is 0.54 versus 0.62 for the market, and the CAPM alpha is –1% per annum, albeit without statistical significance. The category of leveraged and

EXHIBIT 8

Performance and Estimated Exposures of Broad ETF Groups, 2004–2017

ETF Category	All	Market	All ex Market	Lev/Inv	All ex Market ex Lev/Inv
Data start	Jan-04	Jan-04	Jan-04	Jul-06	Jan-04
Return	8.79%	9.23%	8.63%	–5.20%	9.46%
Volatility	13.56%	13.34%	13.84%	9.10%	14.74%
Sharpe ratio	0.56	0.60	0.54	–0.7	0.56
CAPM alpha	–0.75%**	–0.18%	–1.00%	–7.45%**	–0.75%
Market (same period)					
Return	9.70%	9.70%	9.70%	9.97%	9.70%
Volatility	13.75%	13.75%	13.75%	14.72%	13.75%
Sharpe ratio	0.62	0.62	0.62	0.60	0.62
Six-factor alpha	–0.80%**	–0.44%**	–0.89%*	–5.96%**	–0.67%
Mkt-RF	0.97**	1.00**	0.96**	0.10	1.02**
SMB	0.06**	–0.12**	0.16**	0.02	0.17**
HML	0.00	0.01	–0.01	0.27**	–0.02
WML	0.01**	0.00	0.03**	0.07	0.01
RMW	0.00	0.04**	–0.03	–0.13	0.00
CMA	–0.01	0.03**	–0.03	–0.33	–0.02

NOTES: Lev/Inv denotes leveraged and inverse ETFs. **Statistically significant at the 5% level, *Statistically significant at the 10% level.

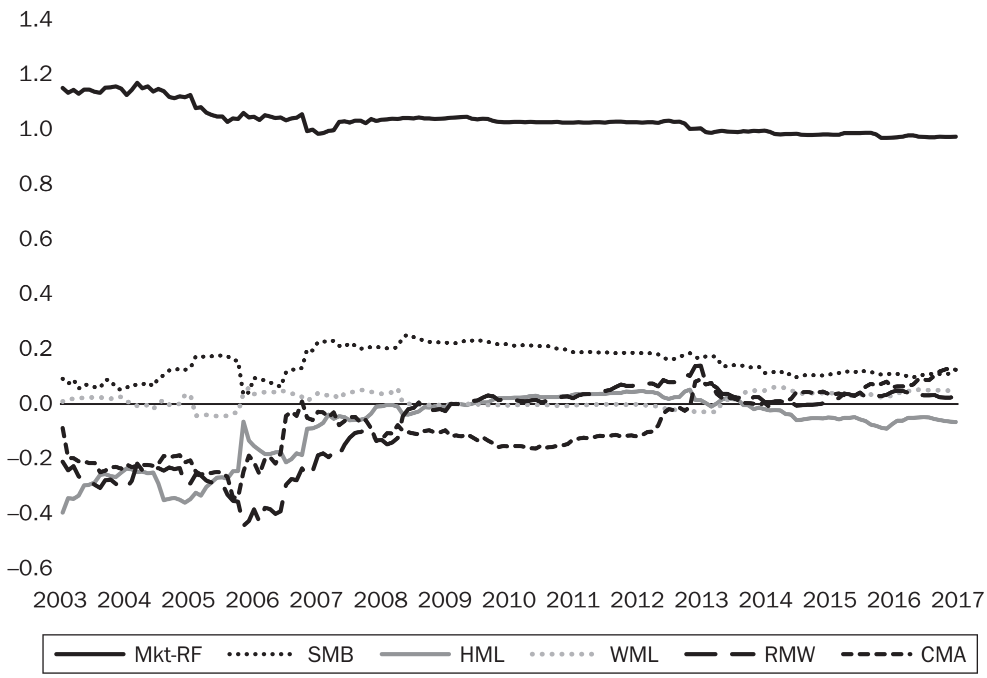
inverse ETFs exhibits particularly dismal performance, with very large and statistically significant negative alphas. Therefore, we also look at the ETFs that neither track a broad market index nor belong to the leveraged and inverse equity category. For this group, we find a slightly lower but still sizable and consistent performance shortfall in terms of raw returns, Sharpe ratios, and CAPM alphas. The negative CAPM alphas are of the same order of magnitude as the alphas reported for the conventional actively managed funds, suggesting that the investors in these ETFs have collectively not been much better off.

Exhibit 9 shows the rolling 60-month factor exposures of the ETFs that neither track a broad market index nor belong to the leveraged or inverse equity category. We observe that their factor exposures have not been constant over time. In addition, pronounced negative factor exposures have been more common than positive ones. The main negative factor exposures are anti-value and anti-low-risk exposures during the early years of the sample. The main positive factor exposure is a small-cap exposure, consistent with a large number of dedicated small-cap ETFs that are present over this period. Toward the end of the sample, the exposures to most factors are, on aggregate, close to zero. This result is in line with Blitz (2017), who examines the factor exposures of US equity ETFs at a recent point in time and also finds that the significant variation that is present at the level of individual funds largely cancels out when all funds are aggregated.

Performance of Factor ETFs

We next show the results for the factor ETFs identified by name. Specifically, we screen for small-cap, value, dividend, momentum, quality and low-risk ETFs. In Exhibit 10, we regress the value-weighted returns of each category on our full set of explanatory factors. These regression results support our name-based classification, as we find that the small-cap ETFs exhibit a high loading on the SMB (size) factor, value ETFs exhibit a high loading on the HML (value) factor, momentum ETFs exhibit

EXHIBIT 9
Rolling 60-Month AUM-Weighted Factor Exposures



a high loading on the WML (momentum) factor, and low-volatility ETFs exhibit a low market beta. The only exception is the quality ETFs, which do not show significant exposures to the Fama and French RMW and CMA (quality) factors. This does not come as a surprise given the finding of Kyosev et al. (2018) that the term "quality" is quite elusive and that quality metrics used by index providers tend to be quite different from the ones used in the academic literature. For the high-dividend funds, we find a mix of high value, profitability, and investment loadings and a low market beta. In some cases, we also observe some pronounced secondary exposures, such as large significant loadings on the size factor (SMB) for the value and momentum categories.

Exhibit 10 also summarizes the performance of each name-based factor category. Small cap is the only category that delivered a higher raw return than the market. This higher return goes along with higher volatility, however, and the small-cap category shows an underperformance on a risk-adjusted basis, with a lower Sharpe ratio than the market and negative alphas. The value, dividend, momentum and quality categories show lower raw returns than the market, lower Sharpe ratios, and also negative alphas. Only the low-risk ETFs show a higher Sharpe ratio than the market and a positive CAPM alpha, but the sample period for this category is quite short, with data starting only mid-2011. This disappointing performance of factor ETFs is in line with Gluskov (2016), who examines the performance of a smaller sample of smart beta ETFs.

We proceed by conducting a similar analysis using the alternative, regression-based classification approach. Specifically, we screen for ETFs with small-cap, value, momentum, profitability, investment, or low market beta exposures—that is, the same factors that we use as control factors when computing the six-factor alphas. At the end of our sample, the average number of funds per factor amounts to about 100, varying between a low of about 55 for momentum and a high of almost 185 for the small-cap factor. Across the board, the regression-based classification approach

EXHIBIT 10

Performance and Estimated Exposures of Factor ETFs Identified by Name, 2004–2017

ETF Category	Small	Value	Dividend	Momentum	Quality	Low Risk
Data start	Jan-04	Jan-04	Jan-04	Jan-04	Jan-06	Jan-11
Return	10.65%	9.51%	8.42%	9.11%	7.98%	12.50%
Volatility	17.86%	15.03%	13.18%	15.52%	15.13%	8.84%
Sharpe ratio	0.53	0.56	0.55	0.51	0.45	1.28
CAPM alpha	–0.88%	–0.76%	0.00%	–0.98%	–1.44%	4.87%**
Market (same period)						
Return	9.70%	9.70%	9.70%	9.70%	9.83%	13.19%
Volatility	13.75%	13.75%	13.75%	13.75%	14.48%	11.40%
Sharpe ratio	0.62	0.62	0.62	0.62	0.60	1.06
Six-factor alpha	–0.53%	–0.61%	–0.83%	–0.73%	–2.78%	1.86%
Mkt-RF	1.01**	0.99**	0.86**	1.03**	1.03**	0.74**
SMB	0.81**	0.15**	–0.10	0.28**	0.12	–0.05
HML	0.09**	0.25**	0.31**	–0.22*	–0.10	–0.21**
WML	0.02**	–0.02	–0.04	0.14**	0.16**	0.24**
RMW	0.02	0.05**	0.25**	–0.17**	0.15	0.28**
CMA	–0.06*	0.07**	0.22**	0.01	–0.09	0.66**

NOTES: **Statistically significant at the 5% level, *Statistically significant at the 10% level.

selects considerably more funds than the name-based approach, which is not surprising because the regression-based approach can also pick up funds that are not primarily designed to give exposure to a certain factor but that do bring along systematic factor tilts, such as sector ETFs. In Exhibit 11, we regress the value-weighted returns of each category on our full set of explanatory factors over their maximum available time span. These results support the regression-based classification approach, as for each category, we find a high and significant exposure to the target factor. Exhibit 11 also summarizes the performance of the smart beta ETF categories identified with the regression approach. Although for some factor categories we observe a higher raw return than for the market, on a risk-adjusted basis, there is little evidence for added value, with Sharpe ratios that are below or similar to the market and alphas that are negative or around zero. Again, the only exception is the low-risk category, which exhibits a higher Sharpe ratio and positive alphas.

Altogether, these results imply that factor ETFs have yet to prove that they can beat the market over prolonged periods of time. In fact, regardless of whether we identify factor ETFs based on name or based on time-series regressions, the realized performance of investors in these ETFs does not appear to be much better than the kind of performance one can expect from conventional actively managed funds. It should be noted, however, that the environment for factor-based approaches has not been favorable, as the performance of classic academic factors over the recent period has fallen well short of the long-term performance realizations; see, for instance, Arnott et al. (2019).

Performance of Anti-Factor ETFs

We next analyze the performance of the ETFs with negative realized factor exposures. Exhibit 12 shows that most groups underperformed both in terms of average returns as well as the Sharpe ratios. Only the growth and aggressive investment ETFs generated higher returns than the market, but they did so

EXHIBIT 11

Performance and Estimated Exposures of Factor ETFs Identified Using Regressions, 2004–2017

ETF Category	Small	Value	Momentum	Profitability	Investment	Low Risk
Data start	Jan-04	Jan-04	Jan-04	Jan-04	Jan-04	Jan-04
Return	10.47%	9.08%	10.25%	9.94%	8.81%	8.68%
Volatility	16.73%	15.96%	15.68%	14.33%	12.69%	11.41%
Sharpe ratio	0.56	0.50	0.58	0.61	0.60	0.66
CAPM alpha	–0.70%	–1.54%	–0.22%	0.09%	0.03%	0.74%
Market (same period)						
Return	9.70%	9.70%	9.70%	9.70%	9.70%	9.70%
Volatility	13.75%	13.75%	13.75%	13.75%	13.75%	13.75%
Sharpe ratio	0.62	0.62	0.62	0.62	0.62	0.62
Six-factor alpha	–0.47%	–1.06%	–0.15%	–0.67%	–0.48%	0.01%
Mkt-RF	1.04**	0.99**	0.99**	0.99**	0.90**	0.86**
SMB	0.52**	0.21**	0.54**	0.18**	–0.05	–0.17**
HML	0.02	0.32**	–0.06**	0.11**	0.15**	0.03
WML	0.00	–0.03	0.08**	0.01	0.00	0.02
RMW	0.02	0.02	–0.03	0.18**	0.13**	0.12**
CMA	–0.08	–0.02	–0.05	0.05	0.20**	0.19**

NOTES: **Statistically significant at the 5% level, *Statistically significant at the 10% level.

EXHIBIT 12

Performance and Estimated Exposures of Anti-Factor ETFs Identified Using Regressions, 2004–2017

ETF Category	Large	Growth	Losers	Unprofitable	Agg. Inv.	High Risk
Data start	Jan-04	Jan-04	Jan-04	Jan-04	Jan-04	Jan-04
Return	8.74%	10.21%	8.57%	8.19%	10.12%	9.19%
Volatility	12.88%	15.10%	15.22%	17.06%	15.37%	16.66%
Sharpe ratio	0.59	0.60	0.49	0.41	0.58	0.48
CAPM alpha	–0.29%	0.09%	–1.90%**	–3.10%**	–0.39%	–2.05%**
Market (same period)						
Return	9.70%	9.70%	9.70%	9.70%	9.70%	9.70%
Volatility	13.75%	13.75%	13.75%	13.75%	13.75%	13.75%
Sharpe ratio	0.62	0.62	0.62	0.62	0.62	0.62
Six-factor alpha	–0.41%	0.46%	–1.51%**	–0.48%	0.08%	–1.42%*
Mkt-RF	0.96**	1.06**	1.04**	1.04**	1.09**	1.08**
SMB	–0.19**	0.12**	0.02	0.18**	0.02	0.34**
HML	0.09**	–0.38**	0.13**	0.03	–0.15**	–0.02
WML	0.00	0.01	–0.05**	–0.03	–0.02	–0.02
RMW	0.02	–0.14**	–0.01	–0.47**	–0.11**	–0.08*
CMA	0.02	–0.05	–0.02	–0.21**	–0.20**	–0.09

NOTES: Agg. Inv. = aggressive investment ETFs. ** Statistically significant at the 5% level, * Statistically significant at the 10% level.

at higher volatility levels, resulting in lower Sharpe ratios. The ETFs with negative momentum (losers), poor profitability, and high risk show particularly poor performance, with statistically significant negative CAPM alphas of –1.90%, –3.10% and –2.05%, respectively.

Exhibit 12 also reports the ex post estimated factor exposures of the anti-factor ETF groups. Mechanically, all groups show large, negative exposures to the factor that we used to classify them. Only in the case of the high-risk ETFs, we see that the market beta is merely modestly high, but that is because this market beta is estimated using the six-factor model; the CAPM beta of this group is 1.18. All groups except for losers have insignificant six-factor alphas, and losers have a negative and significant six-factor alpha despite correcting for their poor exposure to the momentum factor. We conclude that even though the factor ETFs did not generate high returns, in part due to a poor realization of academic factors over our sample, investments in the anti-factor ETFs would have resulted in an even worse performance.

INVESTABLE BENCHMARK MODEL

Out of 23 ETF groups that we consider, 19 have negative six-factor alphas, 4 of which are highly statistically significant. In this section, we aim to distinguish between two possible explanations for the large number of negative six-factor alphas. One potential explanation is that they reflect a significant implementation shortfall of (factor-based) ETFs. An alternative explanation is that the factors used in the six-factor model set an unrealistically high benchmark that is simply not attainable in practice.

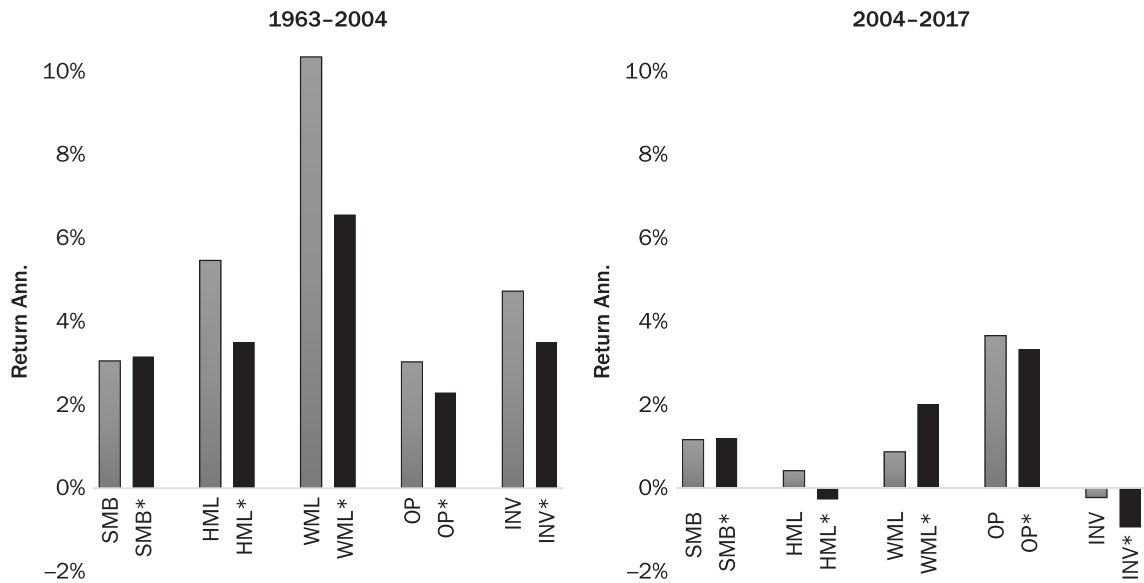
The academic recipe for performance evaluation of investment strategies uses the Fama–French–Carhart (FFC) factors. While this benchmark is widely accepted in the asset-pricing literature, it has certain drawbacks when used for evaluating the real-life performance of investment strategies. Cremers, Petajisto, and Zitzewitz (2012), henceforth CPZ, report that the FFC factors leave economically and statistically significant alphas when used to evaluate the performance of passive benchmarks, such as the S&P 500 and Russell 2000 indices. For instance, they find that over the 1980–2005 period, the Russell 2000 Index had an annualized four-factor alpha of –2.41% (*t*-stat of –3.21). One would not expect this outcome given that the four-factor model contains a broad market (MKT-RF) and a size (SMB) factor and the Russell 2000 index is a straightforward, capitalization-weighted index of small-cap stocks. This turns out to be caused by the ex ante and ex post biases inherent in the FFC factors. For instance, the weight that is placed on the small value stocks that have historically significantly outperformed the small growth stocks in the size (SMB) factor despite being a smaller share of the market. Another example is the disproportional weight of small value compared with large value stocks in the HML factor.

In order to remedy this issue, CPZ propose a modified FFC four-factor model where the portfolios that form the SMB and HML factors are value weighted, as opposed to equal weighted. In this way, the benchmark factors better reflect investable factor strategies. They show that the modified FFC model reduces the index alphas and generates lower tracking errors when used on a sample of actively managed mutual funds.

We construct a modified (investable) FFC six-factor model by value-weighting the 2×3 portfolios that constitute the respective factors. For example, the modified SMB factor is long small value, neutral, and growth portfolios proportional to their market capitalizations and short the large counterparts also proportional to their market capitalization. This effectively means that small value receives less weight in the modified than in the standard SMB factor, as does large value on the short side of the factor. In the case of HML (and analogous for other factors), the portfolio is long large and small value portfolios proportional to their market capitalizations and short the large and small growth portfolios proportional to their market capitalizations. This means that more weight is given to value versus growth performance in the large-cap segment of the market, and less weight to value versus growth performance in the small-cap segment of the market. We also replace the CRSP market factor with the

EXHIBIT 13

Annualized Average Returns of Standard and Modified (*) Factor Premiums



Russell 3000 Index, a capitalization-weighted portfolio that consists of the largest 3,000 US stocks, albeit this change is not consequential as the concerns that CPZ raised about the CRSP market index (inclusion of non-common stocks) no longer hold, because Kenneth French, from whom we source the factor returns, nowadays only includes common shares in the investment universe.

Exhibit 13 shows the annualized returns of the standard and modified (marked with an asterisk) FFC factors over our main sample period (2004–2017) and the pre-sample period (1963–2004). We observe a different pattern over the two samples: In the pre-sample period, we find results consistent with the argument in CPZ that the FFC factors have, in general, generated higher returns than the modified (investable) versions. In our main sample period, however, the performance of the two versions is much closer, and in the case of the momentum factor, the returns of the modified version have been even higher. Therefore, we do not expect our conclusions regarding the performance of ETFs would change substantially if the investable factors were used instead of the standard ones, provided that the coefficients remain comparable. Exhibit 13 also shows that both the standard and the modified versions of the FFC factors have attenuated significantly over the last couple of decades.

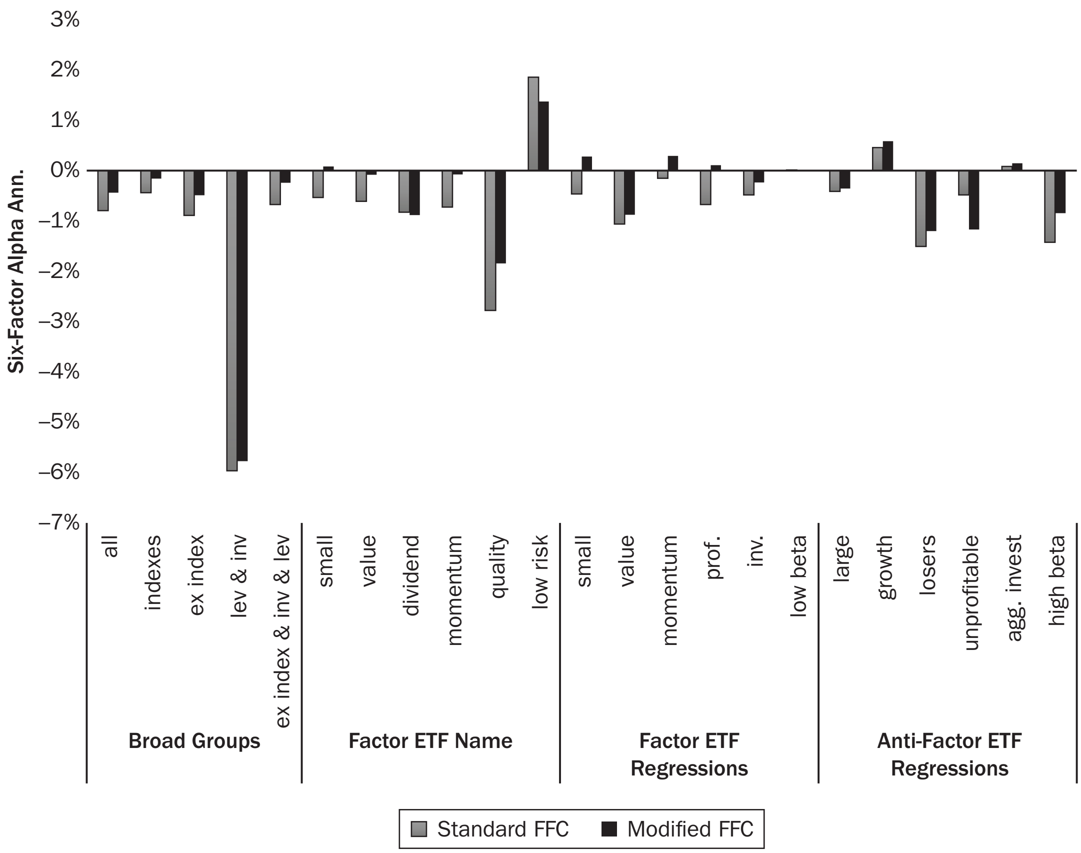
The next step is to conduct the performance evaluation of the broad ETF groups, factor ETFs, and anti-factor ETFs using the modified FFC model. In order to limit the number of exhibits, we only report the six-factor alphas estimated using the standard and the modified FFC models in Exhibit 14 and discuss other results but do not report them in a tabular form in the article.

Across the board, we find that the modified FFC model yields higher R-squared values and lower and more stable tracking errors, consistent with the results of CPZ. When it comes to the broad index trackers, the FFC alpha increases from –0.44% (t-stat. = –1.98) a year to –0.16% (t-stat. = –0.80) a year and loses its statistical significance. In the case of other broad groups (all ETFs, all ex indexes, leveraged/inverse, and all ex indexes & leveraged/inverse), the FFC alphas all get closer to zero and remain significant only in the case of the leveraged/inverse equity group.

Factor ETFs, both the groups identified by name and regressions, have coefficients on the targeted modified FFC factors in line with those of the standard factors, but

EXHIBIT 14

Standard and Modified FFC Six-Factor Alphas



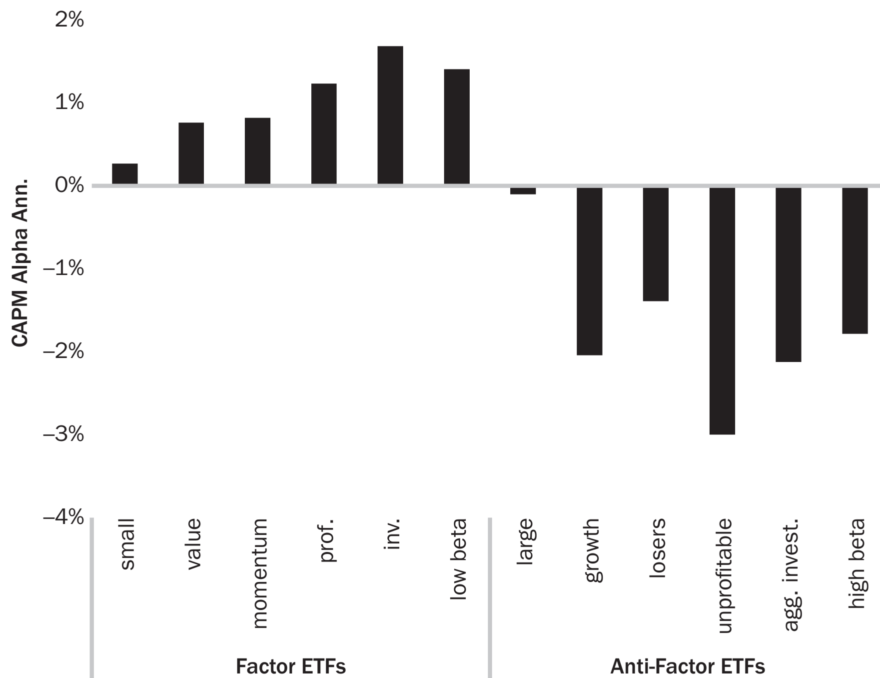
with lower standard errors, yielding higher *t*-statistics and R-squared values, indicative of a better fit. Our conclusions with respect to the FFC alphas remain unchanged: All factor ETFs have alphas that are either negative or close to zero, albeit without statistical significance. In terms of their economic magnitude, the alphas gravitate toward zero in all but two cases (for the dividend group, it remains unchanged, and for the momentum group identified using regression, it changes from –0.15% to 0.29% a year, insignificant from zero in all cases).

When it comes to the anti-factor ETFs, our conclusions again remain unchanged. The modified FFC model provides a better fit, and FFC alphas remain statistically insignificant in all cases but one. The portfolio of past losers continues to generate a significant negative modified FFC alpha of –1.20% (*t*-stat. = –2.22) a year (the standard FFC models yields an alpha of –1.51%, *t*-stat = –2.39). The modified model yields somewhat higher alphas in the case of growth, unprofitable, and aggressive investment groups. For all three groups, however, the alphas remain statistically indistinguishable from zero. Therefore, the leveraged/inverse equity ETFs and the ETFs with negative momentum exposures are the only groups that challenge the ability of the standard and the modified FFC models as performance evaluation tools.

In conclusion, the modified FFC model is able to capture the returns of all but two groups of ETFs in our sample and seems to be a more appropriate benchmark model for ETFs than the standard, asset-pricing FFC model. Given that the sets of factors produce a similar average return over our sample period, however, the main conclusions with respect to the six-factor alphas remain intact.

EXHIBIT 15

Simulated Long-Term Alphas of (anti-) Factor ETFs, 1963–2017



HYPOTHETICAL LONG-TERM PERFORMANCE OF ETFs

The fact that the academic factors had poor returns over our sample period has direct implications for our conclusions regarding the realized performance of the factor and anti-factor ETFs. Given that factor ETFs have provided exposures to their targeted factors, they can potentially be expected to deliver better performance if exposures to these factors are rewarded in line with their longer-term historical realizations. There are, however, other considerations at play as well. First, factor ETFs provide not only an exposure to the intended (targeted) factor or style but also unintended exposures to other factors. Blitz and Vidojevic (2019) emphasize the importance of accounting for both the intended as well as the unintended factor exposures in investment portfolios, given that both affect expected and subsequently realized returns. Second, in almost all cases, ETFs have generated negative returns after accounting for their exposures to common factors—that is, negative six-factor alphas. These alphas, in part, reflect the implementation shortfall of ETFs due to, for instance, fees and costs and would thus depress the return that end investors can expect to reap by investing in these vehicles.

To examine the hypothetical long-term performance of factor ETFs, we take the betas estimated using the six-factor model regressions for each factor ETF group and every month multiply them with realized factor returns to obtain hypothetical return series over the July 1963 to December 2017 period. We specifically include the estimated alphas in this analysis to account for the aforementioned implementation shortfall. Exhibit 15 shows the annualized CAPM alphas of the simulated, long-term factor ETF strategies. All factor ETF portfolios exhibit positive CAPM alphas, albeit they are statistically significant only in the case of profitability, investment, and low risk. This indicates that factor ETFs can be expected to add value if factor premiums manifest themselves as they did in the past.

This exercise also highlights the importance of accounting for all factor exposures. For instance, the low-investment ETFs benefited mostly from a high exposure to the CMA (investment) factor, but they also benefited from high secondary exposures to the value (0.15) and profitability (0.13) factors. At the same time, the negative size exposure was a detractor. In other words, these ETFs do not provide pure exposure to one specific (target) factor but are materially affected by multiple factors. The relatively modest CAPM alphas of the small, value, and momentum factor ETFs can, in part, be attributed to poor exposures to other factors, and in the case of momentum, to a small exposure to the momentum factor itself. In addition, the negative estimated alphas cause a drag on return, with the value ETFs being affected most, given their negative six-factor alpha of 1.06% per annum.

Exhibit 15 also shows the alphas of the simulated anti-factor ETF strategies. Here we see that all groups have negative CAPM alphas, which are statistically significant at the most conservative levels in all cases except for the large category. We also note an asymmetry in performance, where the anti-factor ETFs underperform substantially more than the amount by which the factor ETFs outperform.

CONCLUSIONS

In this article, we analyze a comprehensive, bias-free sample of exchange-traded funds traded on the US exchanges that invest in US equities. We show that the performance of ETFs is not as impressive as one might expect it to be, as investors in these ETFs have collectively realized a performance that does not appear to be much different from the performance that can be expected from conventional actively managed mutual funds. We perform textual and statistical analysis to group ETFs into common investment styles, such as size, value, momentum, quality, and low risk, and show that none of them has managed to consistently add value relative to a capitalization-weighted market portfolio of all US stocks. This can be partly attributed to the generally poor performance of equity factors over much of our recent sample period. Conversely, the anti-factor ETFs, that is, the funds with significantly negative exposures to these factors, have done significantly worse than the market in three out of six cases, and in the other three cases, their performance was not significantly different from that of the market.

Because almost all six-factor alphas are negative, we examine whether this reflects a general implementation shortfall of ETFs or whether the academic Fama–French–Carhart factors pose an unrealistic benchmark. To this end, we reevaluate ETFs against a modified six-factor model with investable factors. We find that this model gives a better fit and leads to alphas that are less negative and gravitate toward zero. Based on these results, we conclude that the standard academic FFC factors lead to conclusions about ETFs' performance that are a bit too negative.

Overall, we conclude that the allure of ETFs finds little empirical support in the data, that ETFs have yet to prove that they can generate better performance than conventional actively managed funds, and that both ETFs and active funds are best evaluated against investable versions of the academic factors that are commonly used in the asset-pricing literature.

ACKNOWLEDGMENTS

The views expressed in this article are solely those of the authors and not necessarily shared by Robeco or its subsidiaries. We thank Matthias Hanauer, Simon Lansdorp, Viorel Roscovan, and Laurens Swinkels for valuable comments on an earlier version of this article.

REFERENCES

Arnott, R., C. Harvey, V. Kalesnik, and J. Linnainmaa. 2019. "Alice's Adventures in Factorland: Three Blunders That Plague Factor Investing." *The Journal of Portfolio Management* 45 (4): 18–36.

Ben-David, I., F. Franzoni, and R. Moussawi. 2017. "Exchange-Traded Funds." *Annual Review of Financial Economics* 9 (November): 169–189.

——. 2018. "Do ETFs Increase Volatility?" *The Journal of Finance* 73 (6): 2471–2535.

Bhattacharya, U., B. Loos, S. Meyer, and A. Hackethal. 2017. "Abusing ETFs." *The Review of Finance* 21 (3): 1217–1250.

Blitz, D. 2017. "Are Exchange-Traded Funds Harvesting Factor Premiums?" *Journal of Investment Consulting* 18 (1): 24–30.

Blitz, D., and M. Vidojevic. 2019. "The Characteristics of Factor Investing." *The Journal of Portfolio Management* 45 (3): 69–86.

Carhart, M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Cheng, M., and A. Madhavan. 2009. "The Dynamics of Leveraged and Inverse-Exchange Traded Funds." *Journal of Investment Management* 7 (4): 43–62.

Cremers, M., A. Petajisto, and E. Zitzewitz. 2012. "Should Benchmark Indices Have Alpha? Revisiting Performance Evaluation." *Critical Finance Review* 2: 1–48.

Easley, D., D. Michayluk, M. O'Hara, and T. Putniņš. 2018. "The Active World of Passive Investing." Working paper, SSRN.

Fama, E., and K. French. 2010 "Luck versus Skill in the Cross-Section of the Mutual Fund Returns." *The Journal of Finance* 65 (5): 1915–1946.

——. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

French, K. 2008. "The Cost of Active Investing." *The Journal of Finance* 63 (4): 1537–1573.

Gluskov, D. 2016. "How Smart Are Smart Beta Exchange-Traded Funds? Analysis of Relative Performance and Factor Exposure." *Journal of Investment Consulting* 17 (1): 50–74.

Kyosev, G., M. Hanauer, J. Huij, and L. Lansdorp. 2018. "Does Earnings Growth Drive the Quality Premium?" SSRN Working Paper No. 2794807.

Petajisto, A. 2016. "Inefficiencies in the Pricing of Exchange-Traded Funds." *Financial Analysts Journal* 73 (1): 24–54.

Robertson, A. 2018. "Passive in Name Only: Delegated Management and 'Index' Investing." Working paper, SSRN.