Financial Analysts Journal | A Publication of **CFA Institute**
https://doi.org/10.1080/0015198X.0000.0000000

Research

Are Passive Funds Really Superior Investments? An Investor Perspective

Edwin J. Elton, Martin J. Gruber, and Andre de Souza

Edwin J. Elton is professor emeritus and scholar in residence at the New York University Leonard N. Stern School of Business, New York. Martin J. Gruber is professor emeritus and scholar in residence at the New York University Leonard N. Stern School of Business, New York. Andre de Souza is assistant professor of finance and economics at the Peter J. Tobin College of Business at St. John's University, Queens, New York.

A number of papers have demonstrated that over historical periods, a specified set of factors has outperformed actively managed funds. In almost all cases, however, the factors used or the procedures followed are not replicable by tradable passive investments. In addition, tradable passive investments have expense ratios that almost always cause them to underperform indexes. The purposes of this article are to identify a small set of exchange-traded funds that captures most of the variation in the population of potential indexes and to determine whether a combination of exchange-traded funds from this small set can be identified that outperforms active mutual funds in future periods.

Disclosure: The authors report no conflicts of interest.

CE Credits: 1

There have been hundreds of papers evaluating the performance of actively managed mutual funds. The vast majority of them have found that after expenses, actively managed mutual funds underperform a benchmark, although the amount of underperformance reported varies among studies.

Given this evidence, academics and investment advisers have recommended that investors use passive mutual funds to meet their investment needs. This recommendation has led to a rapid increase in the proportion of the assets under management that passive funds represent. In the last five years, passive funds have increased from 16.4% of the assets under management to 26%.

This article is motivated by the fact that there is a problem with the methodologies used to make this recommendation, which can affect the size and the existence of the advantage of passive investments. Most of the benchmarks used to evaluate active mutual funds do not represent an investable strategy. It would be impractical to review the hundreds of articles that have been written on mutual fund performance, so we elect instead to review the most common methodologies used in the literature.[1] The most common evaluation methodologies used in mutual fund studies are the Fama–French (2010) three-index model and the Carhart (1997) four-index model. However, there are no available passive portfolios that replicate these indexes. Thus, an investor seeking to use passive portfolios to beat an active fund and attempting to use the Fama–French or Carhart methodology does not have an easily implementable strategy.

A second methodology that is widely used was introduced by Ferson and Schadt (1996). They argued correctly that in evaluating a manager, one should not give the manager credit for following a strategy that is known to produce positive returns—in their case, using past market

We would like to thank Claude Erb and an anonymous reviewer for their helpful comments.

returns to predict future market returns. Although some of their indexes have passive fund counterparts, their evaluation procedure involved changing betas monthly. Once again, this is not an easily implementable strategy for an investor.

Finally, consider the characteristic-based evaluation method of Daniel, Grinblatt, Titman, and Wermers (1997). They divided stocks into 125 groups based on market capitalization, book-to-market ratio, and prior-year return characteristics. They then matched each stock in the active fund portfolio to 1 of the 125 groups and used the resulting benchmarks to evaluate the fund. There is no simple procedure to replicate their procedure using passive funds.

There are two papers that are close in spirit to our analysis but differ significantly in implementation. Berk and van Binsbergen (2015) explained returns with a set of factors based on Vanguard funds. They used an 11-factor model that involved a large amount of short selling of the funds.[2] Because of this requirement, the model cannot be implemented practically. Cremers, Petajisto, and Zitzewitz (2013) showed that the Fama–French and Carhart models do not price standard indexes correctly. They then showed that incorporating a large set of indexes into the model improves performance. Our analysis differs from theirs in that (1) we searched for a parsimonious set of indexes that correctly price other indexes and (2) we show that exchange-traded funds (tradable assets)—rather than indexes—can be used to construct a set of portfolios that outperform active mutual funds.

The purpose of this study is to determine whether a small set of exchange-traded funds (ETFs) can be found that will match the risk of an active fund in one period and outperform that active fund in the following period.

How does our study differ from previous studies? First, we used ETFs in our passive portfolios. These ETFs can be both held long and sold short. Their returns are lower and more variable than the returns on the indexes they follow because of expenses and transaction costs. Although a combination of ETFs that matches the risk (return pattern) of an active fund in any given period can be found, the real issue is not the performance in that period but whether the risk-matching combination in that period will outperform the actively managed fund in the next period. Performance measurement in one period is interesting, but it is much more useful if it can be used to obtain increased performance in future periods.

We will examine risk matching and future performance for two cases: when short sales are allowed and when short sales are not allowed. Throughout our analysis, we will be using easily replicable strategies.

We found that a combination of five ETFs captures most of the variation in all available ETFs. We used these five ETFs to match the risk of active equity funds every December starting in 2004 using two years of previous monthly data. We then compared the return of each active fund with its risk-matching ETF portfolio in the subsequent year. The risk-matching ETFs outperformed the active mutual funds 77% of the time with unlimited short sales and 78% of the time with no short sales. When we considered loads and transaction costs, the percentage of time the ETFs had higher returns rose above 90%. The average extra return varied from 1.37% per year to 1.44% per year, depending on the matching procedure. In addition, the risk-matching ETFs on average had a lower standard deviation of returns.

We examined many of the suggestions for how to select active mutual funds, and although expense ratios, Morningstar ratings, and R^2 criteria increased performance, none of these criteria produced returns above those of the risk-matching ETFs. Finally, we considered the strategy of simply investing in an ETF that matched the active fund's prospectus benchmark. The ETF outperformed the active fund 72% of the time, with an excess return of 1.01% per year. However, the five-ETF model outperformed the prospectus benchmark model at a statistically significant level.

Samples

We used two different samples. The first sample consists of a subset of active equity funds, and the second consists of all ETFs that are candidates for matching the active funds.

The active fund sample consists of all active mutual funds that existed in Morningstar in January 2003 (and met the criteria discussed below). This selection gave us up to 15 years of return history on each fund. We selected this starting date because of the lack of multiple types of ETFs with histories before 2003. To ensure that we had primarily US equity funds, we required all mutual funds to be classified

as US equity funds by Morningstar and to hold at least 90% in equity as of January 2003. In order to deal with incubator bias and bias present in small funds, we eliminated all funds with total assets less than $15 million and those that had existed for less than three years as of January 2003.[3] Finally, we eliminated funds that invested in a specialized sector, levered and long–short funds, asset allocation funds, funds that were used to back variable annuities, and funds that were classified as passive funds. After these eliminations, we were left with 883 funds.

Our second sample (the ETF sample) consists of all ETFs listed in Morningstar. Since our purpose was to match the performance of actively managed funds that invest in a broad spectrum of US equities, we eliminated a number of categories: industry ETFs, real estate ETFs, commodity ETFs, bond ETFs, and international ETFs. Finally, we eliminated actively managed ETFs, leveraged ETFs, and ETFs that follow private indexes.

For the purpose of developing a small set of ETFs to match active funds, we identified the indexes that the remaining ETFs followed. There were 69 unique indexes represented, some of which were followed by more than one ETF. We used these indexes rather than the ETFs themselves to identify a small set of ETFs to use to match active funds. We used indexes rather than the ETFs because for the indexes we had a full history over our 15-year sample period, whereas many ETFs did not exist over our full sample period.[4] Later, when matching the risk of the active funds, we used the ETF with the lowest expense ratio at any point in time for each index selected. Elton, Gruber, and de Souza (2018) showed that selecting the lowest-expense-ratio ETF from all ETFs following the same index is optimum or very close to optimum in all cases.

Determining a Subset of Indexes That Span the Characteristics of All Indexes

In this section, we discuss how we found a small set of indexes and subsequently a small set of ETFs that can explain the return behavior of, respectively, the 69 indexes and a sample of active mutual funds.[5] Many of the 69 indexes are designed to capture return patterns of the same sector of the market (e.g., small growth stocks) and represent only small variations of one another. Our first task was to partition these indexes into groups with similar return patterns and to find the best index to represent each group.

Because the market is a major component of the return on all indexes and because we wanted to capture influences beyond the market, we first removed the CRSP market index return from the return on each index. As discussed in Appendix A, we performed Ward's cluster analysis on 15 years of index returns—with the market index removed—to obtain our groups.

Our cluster analysis started with each of the 69 indexes as its own group and then combined the two indexes whose return patterns were most similar into a group. This process was continued in steps, combining indexes and groups until all were in one group. As discussed in Appendix A, 11 groups seemed to cover all major sectors and subsectors of the market and explained the time-series behavior of all 69 indexes. These groups and the index that best represents each group are shown in **Table 1**. The index that best represented each group was defined as the index with the highest squared correlation with all other indexes in that group.

All 11 indexes may not be necessary. The cluster analysis showed that four groups captured much of the return variation captured by the 11 indexes. These four groups are best described as large value, large growth, small growth, and midcap value. Furthermore, using four indexes plus the market is consistent with much of the literature that has found that five indexes are sufficient to explain the behavior of mutual funds.

Fama and French (2017) provided a testing framework for determining whether a given set of indexes is necessary and sufficient to explain the return pattern of a larger set of indexes. Following their procedure, if the four indexes capture the return variation in the seven additional indexes, then regressing each of those seven indexes on the four selected should result in alphas that are close to zero and not statistically significant. The results are shown in **Table 2**. Using the Wald statistic, we cannot reject the hypothesis that the intercepts, jointly and individually, are zero.

It appears that our 4 indexes are sufficient to price the indexes in the 11-index set. Next, we examined whether four indexes are necessary or whether they could be reduced to a smaller set. Following Fama and French (2017), we regressed each of our four indexes on the other three indexes, all with the

Table 1. Representative Indexes for Each Group

Group	Index	Mean Squared Correlation
Large cap	S&P 500	0.496
Large value[a]	Russell 1000 Value	0.666
Large growth[a]	Russell 1000 Growth	0.643
Midcap	S&P MidCap 400	0.762
Midcap growth	Russell Midcap Growth	0.631
Midcap value[a]	Russell Midcap Value	0.675
Small cap	S&P SmallCap 600	0.783
Small growth[a]	Russell 2000 Growth	0.801
Microcap	Wilshire US Micro	0.811
Momentum	S&P 500 Momentum	0.590
Mixed value	S&P MidCap 400 Pure Value	0.671

Note: This table shows the index with the highest average-in-group squared correlation and the group name for each of the 11 groups we identified.

[a]Indicates inclusion in the four-index model.

Table 2. Non-Base Regressions

Non-Base Index	Intercept	Russell 1000 Growth	Russell 1000 Value	Russell 2000 Growth	Russell Midcap Value	R^2
S&P 500 Momentum	0.067	0.886*	0.152	0.136	−0.076	0.15
Wilshire US Micro	0.057	−1.149**	−0.436	0.737**	0.241	0.59
Russell Midcap Growth	−0.056	0.799**	−0.362	0.287**	0.400**	0.70
S&P 500	0.009	0.484**	0.594**	−0.027**	−0.100**	0.92
S&P MidCap 400	−0.020	−0.075	−0.541**	0.208**	0.735**	0.79
S&P SmallCap 600	0.054	−1.366**	−0.912**	0.464**	0.518**	0.81
S&P MidCap 400 Pure Value	−0.188	−2.353**	−1.856**	−0.338**	1.598**	0.65

Notes: This table shows the results of regressing the representative index from each group not in the base on the four base indexes. The returns on all indexes have the CRSP market index removed: We regressed each index on the market index and used the intercept plus the residual for the returns.

*Significant at the 5% level.

**Significant at the 1% level.

market removed. The results are shown in **Table 3**. Note that in each case, the alphas are much larger than those in Table 2. In addition, each alpha is statistically significantly different from zero at the 1% level, both jointly and individually. We conclude that 4 indexes (along with the market) are sufficient to represent the set of 11 indexes. As explained in Appendix A, we then examined whether replacing the five indexes (the market plus four indexes) with ETFs allows us to explain the return behavior of a small sample of active funds. Since the answer is yes, we turned to examining whether the five ETFs can be used to form a portfolio that will outperform active funds.

The Performance of Active Funds

For each calendar year, using two years of past data, we constructed a portfolio of our five base ETFs that

Table 3. Base Regression Slopes

Base Index	Intercept	Russell 1000 Growth	Russell 1000 Value	Russell 2000 Growth	Russell Midcap Value	R^2
Russell 1000 Growth	–0.053**		–1.058**	–0.207**	0.109**	0.90
Russell 1000 Value	–0.055**	–0.799**		–0.203**	0.156**	0.93
Russell 2000 Growth	–0.239**	–2.731**	–3.542**		0.693**	0.74
Russell Midcap Value	0.224**	1.109**	2.108**	0.535**		0.57

Notes: This table shows the results of regressing each base index on the other three indexes in the base. The returns of each index have the CRSP market index removed: We regressed each index on the market index and used the intercept plus the residual for the returns.

**Significant at the 1% level.

matched the risk of each of the 883 active funds in our sample. The return for each of these portfolios was then computed and compared with the matched active fund for the year following the two-year period in which the portfolio was formed. For example, monthly data for 2002 and 2003 for our five base ETFs were used to construct a risk-matching portfolio for each active fund. The return on each matching portfolio in 2004 was then compared with the return on the corresponding active fund in 2004. In computing the return on the active fund and the risk-matching portfolio for the one-year comparison (e.g., 2004), we excluded the first trading day of January to allow the portfolio of matching ETFs to be formed and the ETFs to be purchased before performance was calculated.

Construction of Risk-Matching Portfolios. In this section, we will discuss how we constructed a portfolio of exchange-traded funds that matched the risk of each active equity fund. For each year starting in 2002, two years of data were used to form a portfolio of the five base ETFs that most closely match the two years of monthly returns on each of the 883 mutual funds in our sample.

In forming a matching portfolio, we examined two cases. The first case follows the type of model used in most of the literature, where all ETFs are allowed to be sold short in unlimited quantities. The weights were determined by a constrained linear regression of a given fund's return on the return of the five base ETFs. In order to determine portfolio weights directly, we followed the usual procedure but constrained the regression coefficients to sum to 1. This is the model typically used in return-based asset allocation studies.

The procedure is basically a constrained regression of the following form:

$$\text{Minimize } \sum e_i^2, \text{ subject to}$$

$$(1)\ \sum_{j=1}^{5} B_{ij} = 1 \text{ and}$$

$$(2)\ R_i = a_i + \sum_{j=1}^{5} B_{ij} R_j + e_i,$$

where

R_i is the return of the ith active fund

R_j is the return of the jth ETF

B_{ij} is the sensitivity of the ith fund to the jth ETF

e_i is the residual for the ith fund

This regression is estimated for each fund i at the end of December, using two years of historical data. Because the sum of the five betas is 1, the betas themselves can be interpreted as portfolio weights.

The weights in the short-sales-allowed case reveal extreme positions within most portfolios. Risk-matching portfolios sometimes result in an ETF being sold short with an unrealistic weight in the portfolio (e.g., –175%) while another is held long with a weight exceeding 100%. Such extreme weights do not seem feasible for real-world investors.

Furthermore, since many investors cannot or will not sell ETFs short, we examined the case of no short sales, which is probably the case of most interest. To compute the no-short-sales-allowed case, we simply added the constraint that B_{ij} is equal to or greater than zero for each of the five base ETFs for each

active fund. Once again, B_{ij} can be interpreted as a portfolio weight.

In each method, we required 24 months of data to be available in the risk-matching years (the fit period), but anywhere from 1 to 12 months of data could be available in the evaluation period. This process was repeated for each year through 2018.[6]

The overall weights from the fit period, as well as the average weight for each of the nine Morningstar categories, are shown in **Table 4**.[7] For the no-short-sales-allowed case, the portfolio weights are generally logical and easy to interpret for each of the Morningstar categories. The Morningstar large value group has the highest percentage in the Russell 1000 Value ETF. The large growth group has the highest weighting on the Russell 1000 Growth ETF, and the large blend category has the highest (and close to equal) percentages in the Russell 1000 Growth and the Russell 1000 Value ETFs. The other Morningstar categories generally follow similar logical patterns.

Note that average results for groups using unconstrained short sales (Panel A) involve large amounts of short selling. For example, matching the risk of funds that Morningstar designates as small value funds involves, on average, short selling 80% of the Russell 1000 Growth ETF while putting more than 100% of the original capital in the CRSP market and over 45% in the Russell 2000 Growth ETF and the

Table 4. Average Regression Parameters

Morningstar Category	Freq.	Intercept	CRSP Market	Russell 1000 Growth	Russell 1000 Value	Russell 2000 Growth	Russell Midcap Value	R^2
A. Short sales allowed								
US Fund Large Value	131	−0.132	0.540	−0.048	0.536	−0.033	0.005	0.93
US Fund Large Blend	176	−0.143	0.586	0.205	0.228	0.001	−0.020	0.94
US Fund Large Growth	213	−0.154	0.403	0.668	−0.140	0.086	−0.017	0.93
US Fund Mid-Cap Value	28	−0.090	0.121	0.038	0.079	0.118	0.645	0.92
US Fund Mid-Cap Blend	37	−0.118	0.262	0.125	−0.130	0.230	0.513	0.92
US Fund Mid-Cap Growth	82	−0.127	0.076	0.510	−0.316	0.373	0.356	0.91
US Fund Small Value	40	−0.043	1.007	−0.800	−0.137	0.471	0.458	0.92
US Fund Small Blend	69	−0.091	0.897	−0.572	−0.245	0.546	0.374	0.93
US Fund Small Growth	107	−0.134	0.477	−0.051	−0.361	0.762	0.173	0.92
All	883	−0.130	0.490	0.161	−0.010	0.221	0.139	0.93
B. No short sales allowed								
US Fund Large Value	131	−0.146	0.094	0.129	0.650	0.041	0.086	0.91
US Fund Large Blend	176	−0.164	0.180	0.340	0.346	0.065	0.069	0.92
US Fund Large Growth	213	−0.151	0.080	0.677	0.081	0.116	0.046	0.91
US Fund Mid-Cap Value	28	−0.063	0.072	0.096	0.147	0.145	0.539	0.91
US Fund Mid-Cap Blend	37	−0.080	0.083	0.174	0.078	0.278	0.387	0.91
US Fund Mid-Cap Growth	82	−0.078	0.049	0.365	0.019	0.421	0.146	0.89
US Fund Small Value	40	−0.007	0.017	0.008	0.122	0.427	0.426	0.89
US Fund Small Blend	69	−0.041	0.029	0.018	0.095	0.542	0.317	0.92
US Fund Small Growth	107	−0.108	0.027	0.082	0.026	0.756	0.110	0.91
All	883	−0.120	0.086	0.306	0.211	0.256	0.142	0.91

Note: This table shows the average parameters of a regression of each active fund on the ETFs following the five indexes indicated.

Russell Midcap Value ETF. For individual funds, the amount of short selling is much more extreme.

How closely do the risk-matching portfolios match returns on active funds in the fit period? Table 4 makes it clear that the R^2 between each of the two types of matching portfolios and the corresponding active fund is quite high. The five-ETF model does an excellent job of explaining the variance of returns on actively managed mutual funds. The R^2 for the matching portfolio goes from 0.93 to 0.91 as we move from short sales allowed to no short sales. The differences are small, but they move in the direction we would anticipate. Adding increasingly severe constraints weakens the relationship in the fit period.

The pattern of portfolio weights in the individual portfolios is very different, though logical, in our two cases. The important question is how this difference affects the ability of the risk-matching portfolio to perform relative to the corresponding active fund in future periods.

Forecast Performance. In this section, we examine whether the passive risk-matching portfolios formed in each two-year period produce higher returns than the active portfolios they mimic in the next year. Note that the next year starts at the close of the first trading day after 1 January of the next year.[8] For each passive portfolio, the weights estimated in any two-year period to mimic the corresponding active fund are multiplied by the return on each ETF in the following year, and the return on the actual fund involved is then subtracted. Results are averaged over the life of each fund, and then the overall averages are computed across each Morningstar category and across all funds. The results are shown in **Table 5**.

Note that the matching portfolio composed only of ETFs outperformed the corresponding active fund in the forecast period by 1.44% per year with short sales allowed and by 1.37% per year with short sales disallowed. Both results are statistically significantly different from zero at the 1% level.

As we would expect, the no-short-sales-allowed case has a smaller difference in performance than the short-sales-allowed case. What is surprising is how little is lost by not allowing short sales. Since short sales are not a consideration for most individual investors, the fact that performance changes very little when short sales are disallowed is encouraging.

For the short-sales-allowed case, the matching portfolio outperformed the active mutual fund in 681 of 883 cases, or 77% of the time. With no short sales, it outperformed in 78% of the cases. Although the no-short-sales constraint is usually binding for individual funds, it makes little difference in the performance of our matching portfolios.

There are two additional influences that can modify our conclusions: loads and transaction costs. One works in favor of the ETF model, and one works against it. If we take loads into account, many of the funds that outperformed the ETFs would be dominated by the ETFs. The percentage of time that the ETFs had higher returns than the active funds increases to 90% in the no-short-sales-allowed case.

The trading costs necessitated by our ETF model could weaken its superior performance. Four of the five ETFs in our base five-ETF model have extremely small bid–ask spreads. Examining a number of days on one exchange revealed a maximum bid–ask spread of 0.04 bp per dollar of price. The one exception was the low-cost ETF matching the Russell 2000 Growth Index. The maximum bid–ask spread we observed for this ETF was less than 0.3 bp per dollar of price.

Take an extreme case. Assume 100% of the Russell 2000 Index is sold in one year and other indexes are purchased. Assume the exact opposite the next year, so that every year 100% of the Russell 2000 Index is traded. In this extreme case, taking loads into consideration and disallowing short sales, there are no active funds that would switch from underperforming the ETF to outperforming the ETF.[9]

Transaction costs for buying and selling a given ETF might well have been higher in prior years. However, what is relevant to investors is their size going forward, and current costs are the best estimate. Thus, transaction costs have little effect on our conclusions.

Not only did the matching passive portfolios have higher mean returns, on average, but they also had lower monthly standard deviations of returns over the life of each active fund. In the no-short-sales-allowed case, the matching ETFs' monthly standard deviations were lower by an average of 0.151, and in the short-sales-allowed case, the standard deviations were lower by 0.103.

Next, consider the performance of our matching portfolio compared with the corresponding active fund for each of the nine Morningstar categories.

Table 5. Monthly Difference in Returns

Morningstar Category	Freq.	ETF Better	Return Diff.	Diff. in St. Dev.
A. Short sales allowed				
US Fund Large Value	131	103%	0.1218%**	−0.0346%
US Fund Large Blend	176	147	0.1357**	−0.0688
US Fund Large Growth	213	178	0.1210**	−0.1601
US Fund Mid-Cap Value	28	21	0.0726**	−0.0804
US Fund Mid-Cap Blend	37	31	0.1487**	−0.0489
US Fund Mid-Cap Growth	82	54	0.0804**	−0.2025
US Fund Small Value	40	28	0.1331**	−0.0776
US Fund Small Blend	69	48	0.1275**	−0.0567
US Fund Small Growth	107	71	0.1138**	−0.1175
All	883	681	0.1201**	−0.1030
B. No short sales allowed				
US Fund Large Value	131	113%	0.1386%**	−0.0058%
US Fund Large Blend	176	152	0.1474**	−0.0140
US Fund Large Growth	213	176	0.1129**	−0.1150
US Fund Mid-Cap Value	28	23	0.0632**	−0.1906
US Fund Mid-Cap Blend	37	30	0.1209**	−0.1059
US Fund Mid-Cap Growth	82	43	0.0176	−0.2828
US Fund Small Value	40	30	0.1652**	−0.4996
US Fund Small Blend	69	52	0.1046**	−0.2940
US Fund Small Growth	107	68	0.1054**	−0.3076
All	883	687	0.1143**	−0.1512

Notes: This table shows the difference in return between the risk-matching ETF portfolio and the corresponding active fund. It also shows the number of times the risk-matching ETF portfolio had higher returns. Differential returns are expressed in percentage per month.

**Significant at the 1% level.

Table 5 shows that the matching portfolio outperformed the active fund for each category under each of our two models. For the short-sales-allowed case, the difference in performance is statistically significant at the 1% level in all categories. For the no-short-sales-allowed case, the difference in performance for eight of the nine Morningstar categories is significant at the 1% level. The group whose performance is not statistically significantly different is midcap growth. The midcap growth and midcap value categories showed the smallest difference in return between the matching portfolio and the active fund in both cases. Also note that for each Morningstar category in each of our two models, the standard deviation of the risk-matching portfolio is smaller than the standard deviation of the active fund.

We have shown in this section that it is possible to form portfolios of a small set of ETFs matching the risk of active funds in one year that will outperform those active funds over the following year. Not only do these risk-matching portfolios produce higher returns, but they also have a lower standard deviation of returns. Given the unrealistic degree of short selling involved in the unlimited-short-sales case, we will not discuss results for this case in the rest of this article.

Before moving on, we should comment on the differential return performance over time. In the

no-short-sales-allowed case, the matching portfolio outperformed the corresponding active fund on average in 12 of 15 years. In the short-sales-allowed case, outperformance occurred in 13 of 15 years. Underperformance occurred around the global financial crisis (2007 and 2009) in both cases.

Additions to Our Model. Although the ETF model performs extremely well, we wanted to see whether there were additional ETFs that could improve the ability of the model to outperform active funds. There were three logical candidates to try. We added each of these separately, producing three six-index models. If these ETFs failed to increase performance, then we would gain additional confidence in the model we have advocated. The first was a momentum ETF. Momentum was added because it has been used as an explanation for performance in the financial economics literature. The second and third were ETFs for the Russell Midcap Growth and Russell 2000 Value indexes. It was logical to try both of these because they were the last to be eliminated in our earlier work on clustering and because the two Morningstar classifications where our model showed the smallest differential performance were the midcap value and midcap growth categories.

In adding momentum, we had to take into consideration that a momentum ETF existed for only the last five years of our sample. To compute returns for the other years, we used the momentum index underlying the ETF minus the average expense ratio for the years that a momentum ETF existed. On average, the momentum index received very little weight: 0.09. The impact on average differential return was a slight increase of 0.3 bp per year in the no-short-sales-allowed case. Adding momentum did not improve results in an economically significant manner.

When we added separately the low-cost ETFs matching the Russell Midcap Growth Index and the Russell 2000 Value Index, there was also little change in results. When we added the value index, the average differential return decreased by 0.4 bp per year in the no-short-sales-allowed case. When we added the midcap growth index, the average differential return improved by 0.8 bp per year. Although the overall results changed very little, there was improvement in the differential return in the midcap growth and midcap blend categories.[10]

None of the three additional indexes we examined made a significant difference in our results.

Selecting Funds That Outperform ETFs

As shown in Table 5, some active funds outperformed a combination of the five ETFs. In this section, we will examine whether any of the standard criteria used to select mutual funds can identify funds that produce returns higher than those of our ETF model. In addition, we will examine whether any of the selection variables are correlated with differential return in the following period. We examine the following selection criteria:

1. Past alpha
2. Past expenses
3. Past turnover
4. Morningstar rating
5. R^2 and R^2 plus alpha

In the first three cases, we sorted the fit period values of each of these variables into deciles and then examined the next period's differential return in each decile. None of these three variables (high alpha, low expenses, and low turnover) produced returns for active funds that were better than those for the passive ETFs in the top decile or any other decile. The only one of these three criteria that produced a significant rank correlation was past expenses. The rank correlation for the no-short-sales-allowed case was –0.87.

When examining Morningstar ratings as ranking criteria, we divided funds into six groups: no rating and five Morningstar ratings. We used the prior year's ratings to compute the next year's differential return. The unrated group had by far the worst differential return, by a significant amount. The group rated by Morningstar as a 5 had the smallest differential return, but the ETFs in that group still outperformed the active funds.

Our final ranking procedure was one suggested by Amihud and Goyenko (2013). They used two ranking procedures: (1) dividing funds into five groups by R^2 and (2) dividing funds into five groups by R^2 and then into five subgroups by alpha (producing a total of 25 groups). They showed that R^2 is highly correlated with active share, as was proposed by Cremers and Petajisto (2009).

When examining these procedures for our constrained short-sales-allowed case and the no-short-sales-allowed case, all groups—including the top

group, the low-R^2 group, and the low-R^2 and high-alpha group—underperformed the ETFs. However, Amihud and Goyenko (2013) put no constraints on the values of the indexes they used. When we allowed unlimited short sales, the active funds in the low-R^2 group outperformed the ETFs, consistent with Amihud and Goyenko's findings.

For the short-sales-allowed case, none of the standard ways of picking active funds outperformed investing in the five ETFs. However, considering expense ratios and Morningstar rankings decreased the underperformance of the active funds. And in the unlimited-short-sales case, the R^2 criteria selected funds that outperformed the five-ETF model.

Comparison with Other Benchmarks

Finally, we wanted to compare our results with a simple strategy that an investor could use. The simplest strategy we can envision assumes the investor buys the lowest-cost ETF that follows the active fund's prospectus benchmark and holds the ETF for the following year. Unfortunately, we could not find a readily available source that provided the prospectus benchmark over past years. Given this constraint, we used the prospectus benchmark the fund used at the time of our study. This method is not perfect, but it is the standard used by other authors. For example, Sensoy (2009) used the current prospectus benchmark for past periods. He justified this method on the basis of conversations with data providers and mutual fund managers. He attributed the fact that prospectus benchmarks change rarely to the fact that the SEC frowns on benchmark changes.

When we assumed the investor selects the ETF that matches the prospectus benchmark, we found that in the period after selection this procedure results in an average excess return of 1.01% per year and a lower standard deviation than that of the active fund. This simple strategy resulted in higher excess returns 72% of the time. That figure is 35 bps less than the value for the five-index model in the no-short-sales-allowed case. This difference is statistically significant at the 1% level.

We tried one other strategy. Active mutual funds need not have a beta of 1 to their prospectus benchmark. Thus, we examined a second procedure to allow for this condition. We allowed the investor to invest in two assets—a riskless asset and the ETF matching the prospectus benchmark—to match the risk of the active fund in every two-year period. We then used these portfolio weights to compute differential returns in the subsequent period. Although this procedure resulted in a higher R^2 in the fit period, the differential return was reduced to 0.81%, and the matching portfolios had higher differential returns 69% of the time in the evaluation period. Thus, in most cases, the simple procedure of buying the lowest-cost ETF that matches the prospectus benchmark is superior to buying the active fund but is inferior to holding the portfolio of ETFs based on the five-index model.

Conclusion

There have been dozens of studies evaluating mutual fund performance after expenses. The general conclusion is that active mutual funds underperform the model used to evaluate their performance. This conclusion has led many investors to switch from active funds to passive funds. The problem is that most models used to perform the evaluation do not represent an investable strategy. The purpose of this study was to examine whether an investor interested in a particular active fund could identify a small set of passive funds that would outperform the active fund in future years. We used exchange-traded funds as our passive funds. They were selected because they can be shorted as well as held long.

Our first task was to find a small set of ETFs that could be used to match the risk of active funds. We initially selected all indexes that are used by any ETF as a benchmark. We then removed the market from each index's return. The market index was reintroduced after a small set of indexes was selected. We then used cluster analysis to obtain 11 groups of indexes in which each member index had returns that were highly correlated with those of other members of the group. For each group, we selected a representative index. The index we selected was the one that was most highly correlated with other indexes in that group. We showed that 4 of these 11 indexes explain the returns on the other indexes. The returns on each of these four indexes plus the market index were then replaced by the returns on the lowest-cost ETF following each of the five indexes each year over our sample period.

To examine whether ETFs do better than active funds, we selected all active equity funds that existed in 2003 (subject to some criteria delineated in the text). For every year, we formed two portfolios that matched each active fund in risk using the four

ETFs discussed previously plus the market ETF: one allowing short sales and one not allowing short sales. We then computed the differential return between the matching portfolio and the active fund in the next year. The return on about 78% of the matching portfolios was higher than the return on the active fund, with the average differential return with no short sales being 1.37% annually. Many of the active funds had loads. If these were taken into account, the matching portfolio beat the active fund about 90% of the time. In addition, the matching portfolios had lower standard deviations. We tried adding some additional indexes to our five-index model: momentum, Russell Midcap Growth, and Russell 2000 Value. Each addition resulted in a deterioration of the results or an improvement by only a tiny and insignificant amount.

We tried several ways of selecting funds that researchers have suggested to see whether active funds that outperformed the portfolio of five ETFs could be identified. Only the low-R^2 criterion of Amihud and Goyenko (2013) resulted in a group of active funds that outperformed the ETFs and only in the case of unlimited short sales. Past expenses and Morningstar ratings provided information but not enough to lead to active funds outperforming the passive matching portfolios. We then examined how an investor would do if he or she followed a simpler strategy that did not involve any computation: selecting the lowest-cost ETF matching the active fund's prospectus benchmark. When we used the fund's prospectus benchmark, the matching ETF had a higher excess return 72% of the time, with an average outperformance of 1.01% annually in the no-short-sales-allowed case. Investors can outperform active funds by buying the lowest-cost ETF that matches each fund's benchmark, but they can do significantly better by using the five-ETF model we developed in this study.

Appendix A

In this appendix, we discuss how we used cluster analysis to place the 69 indexes and then the ETFs into groups such that indexes within any given group had similar return patterns. Cluster analysis has been used by a number of researchers in finance. For example, Brown and Goetzmann (1997) used cluster analysis to group mutual funds with similar return patterns. Elton and Gruber (1971) used cluster analysis to group firms with similar earning patterns. There are many ways to measure similarity. Thus, there are several ways to form groups. We used Ward's criterion.[11]

Ward's cluster analysis starts with each member as its own group and then, step by step, adds members to a group or combines groups until all are in one group. The sequential nature of Ward's criterion and the ability to use it with no a priori theory as to the correct number of clusters make it appealing for this problem. Because the market is so important in determining return patterns and because we wanted to capture factors other than the market, we removed the market return as defined by CRSP from each index's return. This was accomplished by regressing each index on the CRSP market index and clustering on the return of the intercept plus the residual. The next issue was how many groups we needed. There was no definitive way to decide. We used two criteria: the percentage of the variance of the 69 indexes explained by a given number of groups and whether the selected groups included indexes in the same or similar sectors. Eleven groups accounted for 79% of the variance of all 69 indexes. There are three reasons we felt 11 groups were sufficient. First, the increase in R^2 as we increased the number of groups was tiny. Second, examining 12 groups showed that the 2 groups that could be combined to form 11 groups were measuring the same segment of the market. Third, the correlation of the members of the 2 groups that were combined to make 11 groups was extremely high.

The next issue was whether we needed this many groups. Examining the R^2 showed that there was a discontinuity in R^2 at four groups. Four groups accounted for 58% of the variance of the 69 indexes with the market removed, and the increases in R^2 from the 5th through the 11th groups were much smaller and relatively uniform. As discussed in the text, there are standard ways to test whether a certain number of groups is necessary and sufficient to capture the return pattern of all groups. When we performed these tests, four groups were necessary and sufficient. Having decided that four indexes plus the market were sufficient, we had one more concern. We wanted to capture return patterns of active funds with actively traded ETFs, and there may be return patterns in active funds that are not present in the ETFs. Since our goal was to match active investments against a small set of passive investable assets at this stage (and throughout the rest of the study), we replaced each of the four indexes discussed previously as well as the market index with the lowest-cost ETF that follows that index.

To examine whether the five ETFs capture the return pattern of active funds, we selected a stratified sample of 100 active mutual funds that existed over 15 years. The stratification was designed to match the fraction of funds in each of the nine Morningstar categories. Since we insisted on 15 years of data, the sample obviously has survivorship bias. However, we were measuring not performance but, rather, similarity in return patterns, so this bias is unimportant. The forecasting test of our model was not subject to survivorship bias.

We next performed a canonical correlation between our base five ETFs and the 100 funds. Canonical correlation identified the linear combination of the five ETFs that was most highly correlated with a linear combination of the 100 funds in our sample. It then extracted the second orthogonal canonical linear combination that explained the highest correlations after the first canonical correlate had been removed. An *F*-test allowed us to compute the number of canonical correlates necessary to explain the data for all funds, and the number is four (which leads to a correlation of more than 99%). Each ETF in our model was important in explaining one or more of the canonical correlates.

Editor's Note

This article was externally reviewed using our double-blind peer-review process. When the article was accepted for publication, the authors thanked the reviewers in their acknowledgments. Claude Erb was one of the reviewers for this article.

Submitted 30 October 2018

Accepted 8 April 2019 by Stephen J. Brown

Notes

1. See Elton and Gruber (2013) for a more detailed review of the literature.
2. The short selling occurs because the authors did not actually use the Vanguard funds as benchmarks; rather, they used an orthogonalized set of these funds. Each of the 11 factors they used involves both a long and a short position in individual Vanguard funds. These positions can be extreme. For example, it appears that (1) their first factor (the market factor) is 100% long in the Vanguard 500 Index Fund but short more than 100% in the Vanguard Extended Market Index Fund and (2) their third factor, corresponding to a small-cap index, is 100% invested in the Vanguard Small-Cap Index Fund and more than 100% short in the Vanguard Small-Cap Value Index and Extended Market Index funds.
3. See Elton, Gruber, and Blake (2001) and Evans (2010).
4. If the index that proved important at a later stage had been one that did not have a matching ETF over our entire sample, we would have had to change our sample period. Fortunately, this did not occur.
5. The designation of the determinants of return is often referred to as "attribution analysis."
6. In requiring 24 months in the fit period, we did not bias our results. We simply assumed that the investor makes forecasts only in cases when he or she has two years of data at the time of the forecast. Requiring 12 months of data to be available in the evaluation period would bias the results since the investor cannot know how long any fund will survive.
7. Morningstar places each domestic stock fund into one of nine categories, as shown in Table 4. We aggregated our active funds by Morningstar classification and report results by category.
8. The following year is always defined as starting at the end of the first trading day of the following year (returns start on the second trading day) because the fit period ends on 31 December and the investor can use the fit period data to buy ETFs on the first trading day in January.
9. We have not taken brokerage costs into account, but several institutions offer trades of ETFs with zero commission. In the no-short-sales-allowed case and with no loads, there are only eight active funds with differential returns less than 12 bps per year.
10. The average differential return for midcap blend (midcap growth) in the no-short-sales-allowed case increased by 4 (7) bps per year.
11. Ward's criterion minimizes the total within-group variance. We examined the results from other criteria. Ward's produced groupings of indexes following similar sectors and thus conformed to investment logic.

References

Amihud, Yakov, and Rusian Goyenko. 2013. "Mutual Fund's R^2 as Predictor of Performance." *Review of Financial Studies* 26 (3): 667–94.

Berk, Jonathan, and Jules van Binsbergen. 2015. "Measuring Skill in the Mutual Fund Industry." *Journal of Financial Economics* 118 (1): 1–20.

Brown, Stephen, and William Goetzmann. 1997. "Mutual Fund Styles." *Journal of Financial Economics* 43 (3): 373–99.

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

Cremers, Martijn, and Antti Petajisto. 2009. "How Active Is Your Fund Manager? A New Measure That Predicts Performance." *Review of Financial Studies* 22 (9): 3329–65.

Cremers, Martijn, Antti Petajisto, and Eric Zitzewitz. 2013. "Should Benchmark Indexes Have Alpha? Revisiting Performance Evaluation." *Critical Finance Review* 2 (1): 1–48.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers. 1997. "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks." *Journal of Finance* 52 (3): 1035–58.

Elton, Edwin J., and Martin J. Gruber. 1971. "Improved Forecasting through the Design of Homogeneous Groups." *Journal of Business* 44 (4): 432–50.

——. 2013. "Mutual Funds." In *Handbook of the Economics of Finance*, vol. 2B. Edited by George M. Constantinides, Milton Harris, and René M. Stulz. Amsterdam: Elsevier.

Elton, Edwin J., Martin J. Gruber, and Christopher Blake. 2001. "A First Look at the Accuracy of the CRSP Mutual Fund Database and a Comparison of the CRSP and Morningstar Mutual Fund Databases." *Journal of Finance* 56 (6): 2415–30.

Elton, Edwin J., Martin J. Gruber, and Andre de Souza. 2018. "Passive Mutual Funds and ETFs: Performance and Comparison." Working paper, NYU Stern School of Business.

Evans, Richard. 2010. "Mutual Fund Incubation." *Journal of Finance* 65 (4): 1581–611.

Fama, Eugene, and Kenneth French. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65 (5): 1915–47.

——. 2017. "Choosing Factors." Chicago Booth Research Paper No. 16–17.

Ferson, Wayne, and Rudi Schadt. 1996. "Measuring Fund Strategy and Performance in Changing Economic Conditions." *Journal of Finance* 51 (2): 425–61.

Sensoy, Berk. 2009. "Performance Evaluation and Self-Designated Benchmark Indexes in the Mutual Fund Industry." *Journal of Financial Economics* 92 (1): 25–39.