The Review of Financial Studies

Out-of-Sample Performance of Mutual Fund Predictors

Christopher S. Jones
University of Southern California

Haitao Mo
Louisiana State University

We analyze the out-of-sample performance of variables shown to forecast future mutual fund alphas. The degree of predictability, as measured by alpha spreads from quintile sorts or cross-sectional regression slopes, falls by at least half post-sample. These declines appear to be primarily the result of changes in the level of arbitrage activity in the market, with mutual fund competition appearing to play a secondary role. We find no evidence that the declines are the result of data snooping or learning by investors or fund managers. Finally, we show that corporate bond fund performance exhibits similar dependence on measures of bond market arbitrage activity. (*JEL* G11, G12)

Received June 5, 2018; editorial decision June 2, 2019 by Editor Lauren Cohen. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

A central question in the mutual fund literature is whether funds with positive alpha, net of fees and costs, can be distinguished, ex ante, from those with negative alpha. An affirmative answer to the question requires that some variable in the investor's information set be associated with future alphas. To this end, a number of studies have investigated the ability of various theoretically or intuitively motivated variables to predict future fund alphas, and a modest number of variables that appear to do so have now been found. Whether these

---

We are grateful for the guidance of the editor and referee and for many useful comments from Joe Chen, Yong Chen, Timothy Chue, Richard Evans, Wayne Ferson, Juhani Linnainmaa, Oguz Ozbas, Jeff Pontiff, Ravi Sastry, and Nadia Vozlyublennaia. Seminar participants at the University of Southern California, the University of Cincinnati, Louisiana State University, the 2017 Young Scholars Finance Consortium, the 2017 Northern Finance Association meetings, the 2016 Financial Management Association annual meetings, the 2018 Midwest Finance Association meetings, and the 2017 SFS Cavalcade provided very helpful feedback as well. Matti Suominen provided data on hedge fund assets under management; Eric Vogt shared corporate bond liquidity data; and Jerry Hoberg shared code used to measure fund competition. We are indebted to all three. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Christopher Jones, University of Southern California, 701 Exposition Blvd., Los Angeles, CA 90016; telephone: (213) 740-9485. E-mail: christopher.jones@marshall.usc.edu.

The Review of Financial Studies 34 (2021) 149–193

doi:10.1093/rfs/hhaa026 Advance Access publication March 5, 2020

alpha predictors continue to perform outside of their original samples is a question we answer in this paper.

The ability of a variable to predict out of sample may be different from its ability to do so in sample for a number of reasons. Data snooping, perhaps resulting from journals favoring statistically significant results, will lead to an upward bias in in-sample predictive performance, which will naturally decline out of sample. Alternatively, market conditions that differ between the in-sample and out-of-sample periods, such as a change in the amount of capital devoted toward arbitrage activity or a change in the size of the mutual fund industry, may result in a difference in the ability of all investors to find nonzero alphas. Finally, investors or fund managers who learn about return predictors from the publication of academic studies may act in a way that decreases the predictive ability of those variables as their relevance becomes more widely known.

Of these three channels, data snooping is perhaps the one most frequently discussed. While it has long been recognized as a problem in nonexperimental fields (Leamer 1978), the more recent proliferation of empirical finance research has led Harvey, Liu, and Zhu (2016) toward the provocative conclusion that "most claimed research findings in financial economics are likely false." In particular, as Lo and MacKinlay (1990) show, the analysis of characteristic-sorted portfolios creates a particularly high potential for data snooping when the characteristic is selected based on previous analyses of the same data. Supporting this view, Linnainmaa and Roberts (2018) conclude that a number of accounting-based equity market anomalies are likely the result of data snooping, as the performance of such anomaly-based strategies is poor after and, most tellingly, prior to the sample periods analyzed in most of the original studies.

In the second channel, trends or fluctuations in market efficiency or competition cause alpha predictors to vary over time in their effectiveness. As the aggregate amount of arbitrage activity increases, for instance, the alphas on individual assets might be expected to shrink toward zero, with mutual fund alphas shrinking similarly as a result. Chordia, Subrahmanyam, and Tong (2014) investigate how the returns to asset pricing anomalies (momentum, value, accruals, etc.) are related to various measures of arbitrage activity. For a number of anomalies, they find that the average return of a characteristic-sorted spread portfolio is substantially reduced by an increase in short interest, the assets under management of hedge funds, and aggregate share turnover. Jones and Pomorski (2016) analyze the evolution of the positive short-run autocorrelation in market returns identified by Lo and MacKinlay (1988). They showed that its disappearance began in the 1970s, far before the publication of the Lo and MacKinlay paper but coincident with the advent of electronic brokerages and accompanying the rise in trading volume. We follow this work by analyzing how arbitrage activity proxies are related to abnormal fund returns.

The ability of mutual funds to generate alpha may be particularly affected by competition from other funds, who are likely to follow similar strategies and pursue similar mispriced assets. Following the theoretical framework of Pástor and Stambaugh (2012), Pástor, Stambaugh, and Taylor (2015) find strong evidence of negative scale effects at the mutual fund industry level. At the individual fund level, both Wahal and Wang (2011) and Hoberg, Kumar, and Prabhala (2018) show that the performance of a given fund is likely to be worsened by an increase in the number of funds with similar styles of investing. Furthermore, it is intuitive that greater competition could also help wash out poor performers, so that competitive effects could attenuate alphas on both the positive and negative sides.

The final channel we consider is learning, which McLean and Pontiff (2016) find to be an important force behind the out-of-sample decline in equity return anomalies. Although our mutual fund setting differs in terms of how the learning channel would operate, we also consider the possibility that investors or mutual fund managers learn from the academic literature and modify their investment decisions or management styles based on published research. Learning may reduce the ability of some characteristic to predict future alphas by several mechanisms. One is that investors (or their advisors) benefit from published research by becoming better able to identify funds with positive and negative alphas, leading to an increase in flows toward positive alpha funds and away from negative alpha funds. As argued by Berk and Green (2004), declining returns to scale imply that these flows push the alphas of both types of funds toward zero. Alternatively, but with similar effect, fund companies may respond to changes in investor interest by raising the management fees of good funds and lowering the fees of bad funds.

In another version of the learning channel, variables lose their ability to predict future alphas because of manipulation by fund managers. As a hypothetical example, consider the active share measure of Cremers and Petajisto (2009), which is positively related to fund alphas in their sample, a result that has garnered significant attention among mutual fund managers and investors. If investors become aware of this relation and direct their investments toward high active share managers, who are on average skilled, then unskilled managers with low active share may attempt to mimic skilled managers by switching to a high active share strategy, even if they view that new strategy as having performance inferior to their previous one. Over time, the measure becomes more contaminated by this pooling of high and low quality managers, causing active share to lose its effectiveness as an alpha predictor.

Each of these channels has a distinct prediction for how the performance of an alpha predictor should evolve over time. Data-snooping biases should be evident only during the in-sample period and should therefore result in a sudden decrease in alpha spreads both prior to and following the original study's sample period. Learning presumably requires more time to have an impact, perhaps only starting to occur after the original study is published in an academic journal. We would therefore expect alpha spreads to decline gradually following the original study's sample as more investors and fund managers become aware of the value of a predictor. What is similar about both of these explanations is that they ascribe importance to the end of the original study's sample or its publication date. In contrast, changes in mutual fund alphas due to time variation in arbitrage activity or competition among mutual funds should be independent of sample periods or publication dates. Alpha spreads could still be lower out of sample under this explanation, but the reason would simply be that equity mispricing has diminished over time as the result of rising arbitrage activity or competition.

Our first main result is that alpha predictors largely fail, out of sample, to replicate their in-sample success. At least half of the alpha spread generated by predictors proposed in the literature disappears post-sample. In some specifications, the decline is significantly higher. Thus, to a potential mutual fund investor, advice from the academic literature on what mutual funds to hold may be much less useful than advertised.

Presample data are only available for some predictors, but the data we have suggest that presample effects are likely larger than in-sample effects. This finding undermines data mining as a potential explanation for poor out-of-sample performance and implies that the correct explanation is likely to involve a gradual decrease in predictability over time rather than one that occurs right at the start or end dates of the sample periods of the original studies.

Time variation in marketwide arbitrage activity or competition among mutual funds could provide such an explanation, and we find support for both. An increase in the arbitrage activity proxies of Chordia, Subrahmanyam, and Tong (2014) is associated with a significant decrease in alpha predictability. Rising mutual fund industry size also appears to drive down alpha predictability. Measures of competition recently proposed by Hoberg, Kumar, and Prabhala (2018) have some explanatory power as well.

When we include out-of-sample or post-publication effects, arbitrage proxies, and competition measures within the same regression, only arbitrage activity remains robustly significant. These results suggest that post-sample and post-publication effects are artifacts of omitted variables bias, present only when time-varying arbitrage is not accounted for. While not as robust, mutual fund competition, especially the total similarity measure of Hoberg, Kumar, and Prabhala (2018), also appears to drive part of the decline in mutual fund performance predictability.

Since the alpha predictors analyzed in the literature were almost exclusively popularized on the basis of their performance in equity funds, we analyze corporate bond funds to provide another sample that is presumably unaffected by data-snooping biases. Though this asset class differs, some of the same characteristics associated with strong equity fund performance could plausibly translate to bond funds. Confirming this would provide additional evidence against data snooping as the primary explanation of the strong in-sample performance of mutual fund alpha predictors.

Though the set of predictors we are able to analyze is limited by the unavailability of bond fund holdings data, we find strong evidence that most alpha predictors are also effective for corporate bond funds over our entire sample period. This implies that the relations between fund performance and alpha predictors identified in the literature are unlikely to be spurious. Again, however, forecasts implied by the predictors considered appear to depend strongly on the degree of arbitrage activity, in this case measured from the bond market, with higher activity associated with lower alpha spreads.

Our paper relates to several strands of the finance literature. Most clearly, we follow the significant literature on mutual fund performance prediction, which we review in Section 1. Our paper is also closely related to work on the persistence of asset pricing anomalies. In addition to the work of Chordia, Subrahmanyam, and Tong (2014), a number of studies have explored how the expected returns of anomaly strategies have held up since the original papers documenting their existence. Schwert (2003) shows, for example, that the size anomaly largely faded following its discovery in the early 1980s. Jones and Pomorski (2016) analyze several anomalies from the perspective of a Bayesian investor, finding that out-of-sample investment performance is improved by allowing, ex ante, for the possibility that the anomaly return may diminish over time. Finally, in the study most closely related to the current one, McLean and Pontiff (2016) examine the out-of-sample performance of 97 variables shown to predict equity returns and find a substantial decline relative to in sample averages.

Our work is also related to a large literature on data snooping in economics and finance. It has long been understood that specification searches, undertaken either by the individual researcher or collectively by many, may result in violations of the assumptions necessary for the validity of the statistical methods used. A number of solutions have been proposed (Lo and MacKinlay 1990; White 2000; Harvey, Liu, and Zhu 2016) for adjusting inference to take such searching into account, but these methods generally require assumptions about the nature of the search undertaken. Our own approach to dealing with the problem, which is simply to wait until additional data are available, is decidedly low tech, but it has the advantage of being relatively free of assumptions. We believe that our results show that poor out-of-sample performance need not be the result of data snooping if market conditions are changing over time, a conclusion that may have validity in other contexts as well.

Finally, this paper is related to work seeking to characterize the distribution of mutual fund managers. Such work has generally found that the average fund alpha, net of fees and expenses, is negative, but there is substantial disagreement within the literature about the fraction of managers that provide positive alphas. In the Bayesian framework of Jones and Shanken (2005), the bootstrap analysis of Kosowski et al. (2006), and the MLE approach taken by Harvey and Liu (2018), a substantial minority of funds do appear to outperform. In contrast, Fama and French (2010) conclude that funds with positive alphas are highly unusual, while Chen and Ferson (2015) find that they are largely nonexistent. Interestingly, in a similar study, Barras, Scaillet, and Wermers (2010) find a substantial fraction of funds with positive alphas prior to 1996, but find that almost no positive alphas existed by 2006. Our finding that arbitrage activity is associated with declining alpha spreads is consistent with this observation.

1. Mutual Fund Alpha Predictors

We attempt to construct a comprehensive sample of mutual fund predictors by examining journal articles published between 1960 and 2015 with the word "fund" in the title or "mutual fund" in the abstract. We examine articles in six finance journals (*Journal of Finance, Review of Financial Studies, Journal of Financial Economics, Review of Finance, Journal of Financial and Quantitative Analysis*, and *Review of Asset Pricing Studies*) and five general economics journals (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economics and Statistics*, and *Quarterly Journal of Economics*). The resultant list includes all studies that we already knew to contain mutual fund performance predictors.

We examine each paper in that list to find if it contained statistically significant evidence (at the 5% level) that some mutual fund predictor not considered in prior work is related to cross-sectional differences in future mutual fund performance. We include predictors regardless of how that significance is established, whether it is in a regression framework or in portfolio sorts. We do not require any particular method of risk adjustment, and we allow for significance to be established in raw returns, alphas from factor models, or returns adjusted for characteristic exposure. We do require that significance is established for actual fund returns, post fees and costs, which we use in our analysis. It is insufficient for the performance of a predictor to be established only on the basis of stock portfolios formed using reported fund holdings.

A final requirement is that the predictor must be constructible using standard data sets, which we define as CRSP, Compustat, IBES, Thompson-Reuters, and a collection of benchmarks from Standard and Poor's, FTSE Russell, and Barclays.[1] This requirement is imposed solely for reasons of practicality, as a number of papers have proposed performance predictors that rely on proprietary or hand-collected data. Interesting examples of such work include Chen et al. (2013), who hand-check SEC filings for information about fund advisors to understand the role of outsourcing in fund management. Another is Christoffersen and Sarkissian (2009), who find that funds located in larger cities, where private information may be more accessible and knowledge spillovers are more likely, perform better than funds in smaller cities. Although one could possibly obtain the in-sample data from the authors of studies such as these,

[1] If predictors were constructed in the original studies using other data sets, we nevertheless include them as long as we are able to reconstruct them with the data sets in our list.

constructing consistent out-of-sample measures would be prohibitively time consuming.

Most of the predictors we consider can be organized into groups. The first contains various measures of performance based on past fund returns. Within this group, Hendricks et al. (1993) show the first statistically significant evidence that lagged past returns, in their case the return over the past year, forecast future fund returns. Carhart (1997) confirms these results but finds that past alphas from his own 4-factor model are even more useful in forecasting future risk-adjusted performance.

More recent work has proposed methodological improvements to the estimation of mutual fund alphas. Mamaysky et al. (2007) show how backtesting can be used to better identify funds with nonzero alphas. Kacperczyk et al. (2014) propose a skill index that combines both market timing and stock picking abilities and show that it strongly forecasts future fund returns.

Other approaches utilize the returns of other funds or nonbenchmark passive portfolios to improve the estimation of fund alphas. Several papers rely on the intuition of Pástor-Stambaugh (2002) that explains how the performance of portfolios related to a given fund can be, for different reasons, informative about the skill of that fund. Cohen et al. (2005) find that funds whose holdings resemble other funds with strong performance records appear to offer superior performance. Busse and Irvine (2006) implement an estimator from Pástor-Stambaugh (2002) that uses the returns on long-lived passive portfolios to construct more efficient estimators of the alphas of shorter-lived funds, and they find that these estimators lead to improved ability to forecast fund performance. Hunter et al. (2014) show how the performance of a given fund can be assessed more accurately by taking into account the returns on a portfolio of all actively managed mutual funds following the same benchmark.

Yet other papers find information in the extent to which funds deviate from their benchmarks. These deviations can be measured from fund holdings, as in the active share measure of Cremers and Petajisto (2009) or the active weight measure of Doshi et al. (2015). They also can be measured from the R-squared of the fund's returns on one or more benchmarks, as in Amihud and Goyenko (2013). In both cases, funds that deviate more from their benchmarks perform better, suggesting that the greater "conviction" of managers who make larger stock-specific bets is associated with ability. Finally, Simutin (2014) shows that a fund's holdings of cash, an asset that is absent from all equity benchmarks, is significantly related to future performance, which he attributes to the flexibility offered by keeping greater cash positions.

Several performance measures are based on analyses of the portfolio formed on the basis of the most recent fund holdings. Elton et al. (2011), for instance, argue that the alpha of the holdings-based portfolio is a better predictor of future fund returns than the alpha computed from the fund's own returns. The so-called "return gap," proposed by Kacperczyk et al. (2008), is defined as the difference between the actual fund returns and the holdings-based returns. A measure of risk shifting can be obtained, as in Huang et al. (2011), by computing the difference between the volatility of the fund's actual returns and that of the holdings-based portfolio. These studies show that a higher return gap is positively related to future fund returns, while funds that exhibit greater risk shifting perform poorly.

Fund holdings have been used to construct a variety of other portfolio characteristics that have been shown to be related to fund performance. In an early example, Grinblatt et al. (1995) showed that funds holding stocks with high momentum tend to perform relatively well. Chan et al. (2002) used holdings data to characterize funds in terms of the size and book-to-market ratios of their holdings, finding that growth funds tend to offer higher benchmark-adjusted returns. Kacperczyk et al. (2005) find that an industry concentration in fund holdings forecasts high future performance, arguing that fund managers may have informational advantages only in some industries. Kacperczyk and Seru (2007), meanwhile, find that managers whose holdings are more sensitive to changes in public information, proxied by analyst recommendations, tend to perform worse. Gupta-Mukherjee (2014) finds that performance is positively related to the tangibility of the fund's holdings, where tangibility is considered to be high when expenditures on property, plant, and equipment are high relative to expenditures on research and development.

Following Berk and Green (2004), a number of studies are concerned with the impact of fund size on performance. Chen et al. (2004) find that larger funds perform worse, particularly for small-cap funds, while fund family size is positively related to performance.[2] While the interpretation of this evidence is somewhat controversial (see Pástor, Stambaugh, and Taylor 2015) in that it is not clear whether performance is an accurate reflection of managerial skill, the empirical relation between alphas and fund size appears relatively strong. Pollet and Wilson (2008) argue that funds should mitigate the effects of diminishing returns to scale by becoming more diversified, and they show that more diversification is indeed associated with an increase in future fund performance.

Other studies have examined fund flows. Gruber (1996) finds evidence of a so-called "smart money" effect, whereby funds with positive inflows outperform those with outflows, a result given further support by the more comprehensive analysis of Zheng (1999). Lou (2012) shows that this effect, as well as return persistence in stocks and funds, can at least be partially explained by the effects that the funds' flows have on the prices of the stocks they hold.

[2] The CRSP Mutual Funds data set has been significantly updated over time, and fund family data (item mgmt_cd) are not available until December 1999 in the most recent versions of the CRSP Mutual Funds data set. As a result, there are insufficient data to assess the in-sample impact of fund family size, as the original sample ends in December 1999 in the original study of Chen et al. (2004). Hence, we do not include fund family size in our predictor list as our analysis requires the in-sample performance of the predictor. Because of this data limitation, we are similarly forced to exclude three other studies that examine fund families, Massa (2003), Nanda, Wang, and Zheng (2004), and Gaspar, Massa, and Matos (2006), that would otherwise satisfy our criteria for predictor inclusion.

Among other results, he finds that funds holding stocks that are held by funds likely to see future outflows tend to perform relatively poorly.

The final set of studies examines the impact of mutual fund fees and costs, both explicit and implicit. Elton et al. (1993) find that fees are negatively related to future fund performance, a result that was confirmed by Carhart (1997) in a more comprehensive sample. Elton et al. (1993) also found a negative relation between performance and fund turnover, presumably because of the impact of transactions costs.[3] Lastly, Bergstresser et al. (2009) show that broker-sold funds underperform direct-sold funds, which is partially, but not entirely, the result of the high distribution fees charged by broker-sold funds.

In total, 27 different predictors from 25 papers meet our criteria. These 27 predictors represent the starting point of our analysis.

1.1 Mutual fund sample

We obtain mutual fund returns (monthly) and fund characteristics such as expenses, total net assets (TNA), fund portfolio turnovers, and investment styles, from Center for Research in Security Prices Mutual Fund (CRSP MF) database, from November 1961 to January 2015. Fund returns are net of expenses, but not of loads. Quarterly fund equity holdings data from 1980 to 2014 are from Thomson Reuters. When we merge these with CRSP MF, we use MFLINK. We exclude fixed income, international, money market, sector, index, target-date, and balanced funds, focusing on active U.S. equity funds.[4] We subject the fund data to a number of screens to mitigate omission bias (Elton, Gruber, and Blake 2001) and incubation and back-fill bias (Evans 2010). We exclude observations prior to the first offer dates of funds, those for which the names of the funds are missing in the CRSP MF database, and those for which the fund's TNA is below $15 million. To prevent the impact of outliers when holdings data are used, we require a fund to hold at least ten stocks to be eligible in our sample. We combine multiple share classes for each fund, focusing on the TNA-weighted aggregate share class.

[3] The literature reports mixed evidence on the relation between turnover and performance. Elton et al. (1993) and Carhart (1997) find negative relation, whereas Ippolito (1989), Wermers (2000), Edelen, Evans, and Kadlec (2007), and Grinblatt and Titman (1994) find no relation or a positive one. The mixed evidence may be attributable to different samples or different performance measures used. Our own performance measure is the Carhart 4-factor-adjusted alpha of reported net fund return, and we confirm its negative relation with turnover in the sample periods of Elton et al. (1993) and Carhart (1997), using the fund data that is either more comprehensive or comparable to that used in previous studies.

[4] We identify and remove index funds both by CRSP index fund flag and by searching the funds' names with key words "exchange-traded|exchange traded|etf|dfa|index|inde|indx|inx|idx|dow jones|ishare|s&p|s &p|s& p|s & p|500|WILSHIRE|RUSSELL|RUSS|MSCI." and exclude target-date funds by searching the fund names with key words "2055|2050|2045|2040|2035|2030|2025|2020|2015|2010|2005|target." U.S. equity funds are defined as those with policy code CS; Lipper codes EIEI, G, LCCE, LCGE, LCVE, MCCE, MCGE, MCVE, MLCE, MLGE, MLVE, SCCE, SCGE, SCVE, CA, EI, G, GI, MC, MR, or SG; Strategic Insight codes AGG, GMC, GRI, GRO, ING, or SCG; or Wiesenberger codes G, GCI, IEQ, LTG, MCG, or SCG. Using the recently introduced CRSP style code (item crsp_obj_cd) instead results in almost the same sample.

Our final sample, which ranges from November 1961 to January 2015, comprises 3,069 unique funds, and the average number of months for a fund in our sample is 147.

1.2 Computing alpha spreads

The ultimate object of interest in our study is mutual fund alpha.[5] To control for standard risk exposures and to allow alphas to vary over time, we follow Carhart (1997) by computing alphas based on rolling window estimates of factor betas. Specifically, for each fund at each date, we use the previous 36 months to estimate the betas on the Fama and French (1993) and Carhart factors. We then use those betas to risk-adjust the current month's excess return. We refer to the result as the "realized alpha," denoted for fund i and time t as α_{it}.

A large part of our analysis is focused on the behavior of spreads in mutual fund alphas. These are formed by measuring the relation between fund alphas and some fund-level predictor x_{ijt}, where the i subscript denotes the fund and j the predictor. For compactness of notation we specify the date as t but note that the predictor is always known as of the end of month $t-1$. In all cases, the predictor variable is defined such that high values are associated with good performance and low values with bad performance, where this determination is made on the basis of the original paper in which the predictor was proposed.

We measure alpha spreads in two ways, by sorting and by cross-sectional regression. Sort-based alphas are computed by sorting on x_{ijt} and computing the difference between the equal weighted average alphas in the top and bottom quintiles. For shorthand we denote this spread as "Q5-Q1." We also use cross-sectional regression to produce alpha spreads. In this approach, we simply run univariate monthly regressions of ex post fund alphas on individual predictors that are normalized to have zero mean and unit standard deviation in the cross-section. The slope coefficients of these regressions constitute the "CSR" alpha spreads.

Two of the predictors that we include (the broker-sold indicator of Bergstresser, Chalmers, and Tufano 2009 and the holdings-based growth fund indicator of Chan, Chen, and Lakonishok 2002) are indicators, always taking values of 0 or 1. For these predictors, the Q5-Q1 spread is computed as the difference between the equal weighted average alphas in the two portfolios, while the CSR spread is computed by running cross-sectional regressions on these indicators. In these cases, these two procedures result in numerically identical values.

Normalization of the predictors in regression analysis is necessary given that we will be using the resultant CSR coefficients in panel regressions. Without normalization, differences in the variances of different predictors will cause large deviations in CSR coefficients that are unrelated to the strength of the

[5] Unless specified otherwise, fund alpha in this paper always refers to the alpha of the actual fund return, net of fees and costs.

predictions. Normalization of the predictors has the effect of making the CSR coefficients from different predictors reasonably homogeneous in terms of time series means and variances.

We believe that including both types of alpha spreads is potentially important. Cross-sectional regression maximizes dispersion in the predictor, which potentially allows funds with very large or very small values of the predictor to have more influence on the estimated spread. This is appropriate if we believe that those funds are more heavily affected by whatever force underlies the effectiveness of that predictor. This is undesirable, however, if the relation between predictor and alpha is nonlinear. In this case, computing spreads based on extreme quintiles makes more sense. Given no guidance to choose one approach or the other, we include them both.

1.3 Criteria for including predictors

As discussed at the beginning of this section, we have identified 27 predictors found in the mutual fund literature to forecast future fund performance in a statistically significant manner using a small number of standard databases. Our goal is to understand the out-of-sample performance of these predictors, so we must first document in-sample performance. We follow McLean and Pontiff (2016) in requiring that predictors exhibit a degree of success in-sample that falls somewhat short of statistical significance, namely, that the average in-sample alpha spread has a *t*-statistic above 1.5. For the extreme quintile (Q5-Q1) results, this *t*-statistic is computed on the basis of in-sample Q5-Q1 spreads. For the cross-sectional regression (CSR) results, the *t*-statistic is based on CSR spreads.

We emphasize that we are not attempting to replicate earlier results but rather to re-evaluate these results within a consistent framework. The mutual fund literature contains many different approaches for assessing the validity of a predictor. These include the comparison of portfolios based on some sorting variable, where the number of portfolios is typically between 3 and 10, and where each portfolio may be equally weighted or value weighted. Regression-based approaches are also common and may use the panel or Fama-Macbeth frameworks, with or without adjustment for serial correlation and heteroscedasticity. In addition, regressions may include control variables, which we do not include in our own analysis. Finally, adjustment for benchmark or factor risk is handled very differently across different papers, with the early literature, in some cases, focusing on raw returns and most papers since Carhart (1997) using the 4-factor model from that paper.

Table 1 shows the results of this analysis of each of the 27 predictors we consider.[6] Using the sample period from the original papers, 18 of the 27

[6] To ease a comparison with original studies, we do not align predictors in Table 1 to ensure that higher values are associated with better future fund performance. In subsequent tables we change predictor sign, when necessary, so that higher values imply better performance according to the original studies.

Table 1
Alpha spreads in and out of sample

			Q5 - Q1				CSR			
Predictor	Sample period	Publication date	In-sample spread	Out-of-sample spread	Post-publication spread	Agree w/ orig?	In-sample spread	Out-of-sample spread	Post-publication spread	Agree w/orig?
Abnormal cash holdings (Simutin 2014)	1992–2009	2014:07	11.84 (2.40)	−3.29 (−1.14)	−14.86 (−1.19)	Yes	2.16 (1.50)	−1.17 (−1.56)	−3.71 (−1.19)	Yes
Active share (Cremers and Petajisto 2009)	1990–2003	2009:09	7.26 (1.02)	1.54 (0.35)	−0.21 (−0.04)	Sign	1.50 (0.56)	1.05 (0.66)	−0.04 (−0.02)	Sign
Active weight (Doshi et al. 2015)	1980–2013	2015:12	7.95 (2.86)	−0.33 (−0.04)		Yes	2.69 (2.86)	−0.12 (−0.05)		Yes
Direct-sold (Bergstresser et al. 2009)	1996–2004	2009:10	4.78 (2.79)	−1.85 (−0.44)	5.16 (3.34)	Yes	4.78 (2.79)	−1.51 (−0.44)	5.16 (3.34)	Yes
Flow-induced trading (Lou 2012)	1980–2006	2012:12	21.86 (2.86)	−7.11 (−0.81)	7.73 (0.63)	Yes	8.22 (3.01)	−2.44 (−0.87)	2.10 (0.54)	Yes
Holdings-based alpha (Elton et al. 2011)	1994–2005	2011:04	20.32 (2.13)	14.97 (2.54)	4.11 (0.50)	Yes	8.24 (2.34)	5.40 (2.49)	2.02 (0.74)	Yes
Industry concentration (Kacperczyk et al. 2005)	1984–1999	2005:08	2.33 (0.40)	0.55 (0.10)	−2.48 (−0.42)	Sign	−1.39 (−0.67)	−0.47 (−0.23)	−1.01 (−0.59)	No
Intangibles (Gupta-Mukherjee 2014)	1980–2009	2014:02	−6.94 (−0.87)	−9.33 (−1.04)	−22.87 (−0.92)	Sign	−3.30 (−1.22)	−2.21 (−1.01)	−3.97 (−0.69)	Sign
Inverse of diversification (Pollet and Wilson 2008)	1976–2001	2008:12	1.49 (0.34)	−1.83 (−0.75)	−3.53 (−1.08)	No	−0.93 (−0.60)	−0.22 (−0.25)	−0.61 (−0.52)	Sign
Fund size (Chen et al. 2004)	1962–1999	2004:12	−4.43 (−0.99)	−0.22 (−0.10)	1.62 (0.96)	Sign	−2.27 (−1.55)	0.05 (0.07)	0.82 (1.39)	Sign
One-year return (Hendricks et al. 1993)	1974–1988	1993:03	33.87 (4.05)	28.57 (3.97)	18.25 (2.31)	Yes	13.12 (4.12)	10.74 (4.20)	6.59 (2.34)	Yes
Momentum (Grinblatt et al. 1995)	1974–1984	1995:12	−13.67 (−1.44)	−1.91 (−0.61)	−4.49 (−1.23)	No	−3.64 (−0.96)	0.12 (0.11)	−1.50 (−1.20)	No
Fund flows (Zheng 1999)	1970–1993	1999:06	8.28 (1.76)	5.34 (1.73)	5.26 (1.39)	Yes	2.28 (1.52)	1.00 (1.26)	0.67 (0.71)	Yes
Public info (Kacperczyk and Seru 2007)	1993–2002	2007:04	1.01 (0.20)	0.39 (0.20)	0.90 (0.35)	No	2.16 (1.24)	0.34 (0.50)	1.08 (1.30)	No

(Continued)

Table 1
(Continued)

			Q5 - Q1				CSR			
Predictor	Sample period	Publication date	In-sample spread	Out-of-sample spread	Post-publication spread	Agree w/ orig?	In-sample spread	Out-of-sample spread	Post-publication spread	Agree w/orig?
Risk shifting (Huang et al. 2011)	1980–2009	2011:08	5.90 (1.00)	−0.80 (−0.11)	3.21 (0.37)	No	0.33 (0.18)	−1.00 (−0.48)	0.10 (0.04)	No
Growth style (Chan et al. 2002)	1985–1997	2002:10	−7.69 (−0.83)	−23.02 (−1.95)	−6.10 (−0.56)	No	−7.69 (−0.83)	−23.02 (−1.95)	−6.10 (−0.56)	No
Skill index (Kacperczyk et al. 2014)	1980–2005	2014:08	23.69 (3.38)	1.54 (0.24)	3.93 (0.27)	Yes	9.34 (3.74)	1.03 (0.47)	3.46 (0.64)	Yes
R-squared (Amihud and Goyenko 2013)	1989–2010	2013:03	−8.20 (−1.81)	7.96 (1.12)	15.26 (1.31)	Yes	−3.21 (−2.07)	2.50 (1.18)	4.64 (1.28)	Yes
One-year Carhart alpha (Carhart 1997)	1962–1993	1997:03	24.84 (3.33)	18.53 (3.61)	18.79 (3.19)	Yes	9.75 (3.30)	6.80 (3.67)	7.21 (3.40)	Yes
Five-year Carhart alpha (Carhart 1997)	1962–1993	1997:03	41.80 (6.98)	13.73 (3.37)	10.58 (2.26)	Yes	15.70 (6.99)	5.11 (3.59)	4.26 (2.61)	Yes
Back-tested alpha (Mamaysky et al. 2007)	1970–2002	2007:09	29.85 (4.84)	27.65 (3.02)	12.28 (2.15)	Yes	10.12 (4.60)	8.55 (2.99)	3.80 (1.75)	Yes
Active peer benchmarks (Hunter et al. 2014)	1983–2010	2014:04	33.34 (6.62)	16.64 (2.88)	15.21 (0.90)	Yes	11.29 (6.68)	5.83 (2.86)	5.97 (1.00)	Yes
Pastor-Stambaugh alpha (Busse and Irvine 2006)	1985–1995	2006:10	34.63 (4.65)	35.77 (7.70)	10.49 (1.66)	Yes	14.37 (4.58)	13.04 (7.57)	3.57 (1.58)	Yes
Success overlap (Cohen et al. 2005)	1982–2002	2005:06	21.97 (2.43)	9.87 (1.57)	−3.03 (−0.42)	Yes	7.85 (2.37)	4.35 (2.12)	−1.39 (−0.61)	Yes
Expense ratio (Elton et al. 1993)	1965–1984	1993:01	−28.69 (−3.70)	−9.23 (−3.81)	−8.71 (−3.15)	Yes	−8.78 (−3.26)	−3.33 (−3.98)	−3.04 (−3.21)	Yes
Return gap (Kacperczyk et al. 2008)	1984–2003	2008:11	18.07 (4.78)	8.02 (2.29)	2.69 (1.02)	Yes	8.54 (6.51)	3.20 (2.52)	1.00 (1.10)	Yes
Turnover (Elton et al. 1993)	1965–1984	1993:01	−19.93 (−1.98)	−6.36 (−1.35)	−8.49 (−1.54)	Yes	−2.39 (−0.65)	−2.04 (−1.44)	−2.33 (−1.44)	Sign

This table reports summary statistics on the alpha spreads, in basis points per month, associated with abnormal return predictors documented in the mutual fund literature. The realized alpha for each fund on each date is computed by subtracting the exposures to the four Carhart factors, where the betas corresponding to each fund/date are computed based on the prior 36 months of returns. Alpha spreads are computed either as the difference between the top and bottom quintiles of a predictor-based sort (Q5-Q1) or as the slope coefficient of a cross-sectional regression of realized alphas on the predictor (CSR). We examine these spreads separately over various subsamples. The table also reports the sample period used in the original study as well as that study's publication date. Finally, the table reports whether our in-sample average spread is consistent with the original study ("Yes"), consistent with the same sign but with a t-statistic below 1.5 ("Sign"), or has the opposite sign as the original study ("No"). The full sample period starts in 1961:11 and ends in 2015:01. t-statistics are in parentheses.

predictors have t-statistics above 1.5, with estimated coefficients that match those in the original studies, for the results based on quintiles (Q5-Q1). Of the remaining nine predictors, four match the sign but fail to meet the 1.5 t-statistic cutoff, and five are estimated with insignificance and with signs that are opposite those in the original studies. For the results based on cross-sectional regression (CSR), 18 predictors are again significant with matching sign, though they are not all the same as those significant under the sort-based analysis. Five of the remaining nine have the matching sign but are not significant, whereas four are insignificant with the opposite sign.

In the cases in which our results are inconsistent with the original studies, the explanation for the inconsistency is sometimes obvious. Grinblatt et al. (1995), for instance, document the effects of momentum on mutual fund returns, establishing that momentum is a significant factor for explaining fund performance. This result was one motivation behind the 4-factor model subsequently proposed by Carhart (1997). Because our risk adjustment is based on Carhart's model, the predictor proposed by Grinblatt et al. (1995) becomes insignificant.

In other cases, our finding of no statistical significance does not contradict the results of the original paper. Cremers and Petajisto (2009), for instance, document the statistically significance of their active share variable, but they show that it is not robust to the factor adjustment that we focus on in this paper. Kacperczyk and Seru (2007) document a strong relation between the reliance on public information and fund alphas, but they only consider regressions in which a number of control variables are included.

In each of these cases, the methodological choices made in the original papers could well be appropriate. Our own simple and standardized framework is likely not the ideal way to evaluate every predictor. Given the impracticality of using a different empirical approach for each predictor and our desire for consistency, it appears inevitable that some predictors will fail to make the significance cutoff.

Table 1 also shows the performance of the predictor-level alpha spreads in data outside the sample periods used in the original study. In the vast majority of cases, alpha spreads that are significant in sample are lower in the out-of-sample and post-publication periods. This motivates the more systematic analysis in the next section.

2. Out-of-Sample Performance

The central question we answer in this paper is whether mutual fund alpha predictors continue to work out of sample. We address this question in several ways. We begin by following McLean and Pontiff (2016) by analyzing how alpha spreads change outside the original sample period. We then consider a related framework for analyzing predictability in a panel of mutual funds.

2.1 Predictor-level descriptive statistics

Our analysis begins by analyzing the panel of predictor-level alpha spreads, with one observation for each predictor in each month, where the alpha spread is defined either as the Q5-Q1 spread or the CSR coefficient. Table 2 describes the data that will be used in our predictor-level panel regressions, which combines data on all predictors shown to be significant and with a sign matching that of the original study.

The table shows that of the 18 predictors included in our Q5-Q1 results, there is a median of 246 months of data in their respective sample periods and 216.5 months of data outside them. Most out-of-sample data follow the end of the original studies' sample periods, but for seven of the 18 predictors we have some data that predates the samples originally used. When presample data exist, they average 108 monthly observations. In contrast, post-sample data average 139 months and are available for all predictors. For the 18 predictors that are significant in the CSR-based analysis, the corresponding figures are similar.

Both Q5-Q1 and CSR spreads typically show sizable in-sample averages. For Q5-Q1 spreads, the median across all 18 predictors is 21.91 basis points (bps) per month, and the median t-statistic is 3.1. The interpretation of CSR-based alpha spreads is not as straightforward. Recall that the CSR-based alpha spread is the slope of a cross-sectional regression in which the predictor has been normalized to have zero mean and unit standard deviation. The average CSR alpha spread is the average coefficient on these normalized predictors. The 8.39 in-sample value implies, therefore, that a 1-standard-deviation increase in a predictor would increase the fund alpha by 8.39 bps per month. The median t-statistic in this case is 3.14.[7]

Out-of-sample spreads are much lower. For Q5-Q1 spreads, the median spread drops to about 8.6 bps per month, with a median t-statistic of 2.01. For CSR, the median spread drops from 8.4 to just 3.3, with a median t-statistic of 2.31.

The table also shows the median value of the ratio of the out-of-sample and in-sample alpha spreads. For the Q5-Q1 case, the median ratio is 0.39, meaning that the typical predictor's out-of-sample alpha is just 39% of its in-sample alpha. For the CSR case, the median ratio is similar, at 0.38. For all predictors in the CSR results, the out-of-sample spread is lower than the in-sample spread. This relation is reversed only once for the Q5-Q1 results, which occurs for the Pástor-Stambaugh (2002) alpha, as implemented by Busse and Irvine (2006), which is slightly stronger outside the original sample period of Busse and Irvine.

[7] Note that the smaller magnitude of the CSR-based alpha spreads is expected. For a normally distributed predictor, the average predictor in the fifth quintile would be about 1.4 standard deviations above the mean and about 2.8 standard deviations above the average value in the first quintile. A difference of 2.8 standard deviations corresponds to an in-sample alpha that is higher by $2.8 \times 8.39 = 23.49$, which is close to the value of 21.91 obtained in the Q5-Q1 analysis.

Table 2
Alpha spread summary statistics

	In sample	Out of sample	Pre-sample	Post-sample	Post-publication	Post-sample pre-publication	In sample pre-1992 only	In sample 1992–2015 only	In sample style-adjusted	Out of sample style-adjusted
A. Q5 - Q1										
Number of predictors included	18	18	7	18	14	18	15	15	17	17
Median number of months	246	216.5	108	139	107	57.5	144	144	169	169
Median alpha spread (bps/month)	21.91	8.63	35.30	5.50	8.11	6.35	26.49	19.83	17.64	8.25
Median t-statistic	3.10	2.01	2.23	1.87	1.60	0.89	3.13	2.40	3.03	1.66
Median ratio relative to in-sample values		0.39	1.52	0.32	0.33	0.29	1.04	0.94	0.85	0.32
Percentage < 1		94	14	100	93	94	20	60	71	100
B. CSR										
Number of predictors included	18	18	7	18	14	18	15	16	17	17
Median number of months	257.5	180.5	108	139	107	57.5	144	138	240	169
Median alpha spread (bps/month)	8.39	3.26	16.11	1.94	2.57	2.01	9.70	6.37	6.97	3.07
Median t-statistic	3.14	2.31	2.68	1.46	1.34	1.22	3.11	2.34	3.14	1.64
Median ratio relative to in-sample values		0.38	1.38	0.32	0.26	0.38	1.06	0.91	0.86	0.35
Percentage < 1		100	14	100	93	94	13	69	76	100

This table reports summary statistics on the alpha spreads associated with various abnormal return predictors documented in the mutual fund literature. The realized alpha for each fund on each date is computed by subtracting the exposures to the four Carhart factors, where the betas corresponding to each fund/date are computed based on the prior 36 months of returns. Alpha spreads are computed either as the difference between the top and bottom quintiles of a predictor-based sort (Q5-Q1) or as the slope coefficient of a cross-sectional regression of realized alphas on the predictor (CSR). We examine these spreads separately over various subsamples. The final two columns of the table compute spreads based on predictors that are normalized by their style mean and standard deviation. The full sample period starts in 1961:11 and ends in 2015:01, except for the style-adjusted results, which start in 1980:1 because of the availability of holdings data.

While presample data are spotty, we separately examine the presample and post-sample periods to determine in which part of the out-of-sample period alpha spreads decline most. Interestingly, we find that Q5-Q1 presample alpha spreads are substantially larger, at around 1.52 times their in-sample values, for the seven predictors that can be analyzed. Presample CSR spreads are only slightly smaller. Thus, the out-of-sample decline in alpha spreads appears to come entirely from the post-sample period.

Following McLean and Pontiff (2016), we also examine the post-publication performance of the alpha predictors and separately consider the portion of the post-sample period that predates publication. While post-publication data is only available for 14 predictors, both for the Q5-Q1 and CSR cases, the performance over these samples is typically similar to post-sample values. In the window between sample end and publication, alpha spreads are lower than the post-publication values, though these estimates should be treated with caution given the relatively short lengths of these periods (median of 57.5 months).

Table 2 also sheds some light on how alpha spreads evolve *within* the in-sample periods. For each predictor with available data, we examine the average alpha spread separately before and after the beginning of 1992, a breakpoint before and after which the median predictor has about the same number of in-sample months. Both for the Q5-Q1 and CSR results, the median predictor's alpha spread over the pre-1992 part of the in-sample period is slightly higher than the full in-sample median, while the median post-1992 in-sample spread is lower. This result suggests that alpha spreads decay during the sample periods of most studies, and not solely after those sample periods are over.

Finally, the table shows how the average in-sample and out-of-sample alpha spreads are affected by style adjustment. As described in more detail in the Online Appendix, we classify funds into a two-by-two grid of styles based on holdings-weighted average percentiles in market capitalizations and book-to-market ratios. For each date, we demean a fund's predictors by subtracting the contemporaneous averages of all funds in the same style and then rescale them using the contemporaneous standard deviations from the same group of funds. Using these style-adjusted predictors in place of the unadjusted predictors produces the results in the last two columns of the table.

Because style adjustment requires mutual fund holdings data, which are only available starting in 1980, the median in-sample and out-of-sample periods are somewhat shorter. Nevertheless, style adjustment appears to have little impact on any of our main results, suggesting that the power of our mutual fund performance predictors is largely independent of investment style.

2.2 Predictor panel regressions

To assess the statistical significance of some of the patterns present in Table 2, we follow McLean and Pontiff (2016) by reassessing the out-of-sample performance of mutual fund predictors in a panel regression setting in which all alpha predictors are analyzed jointly. In this panel we can control for cross-sectional dependence by clustering and serial dependence either by clustering or through the use of predictor fixed effects. Further, panel regression naturally takes into account that our predictors are observed over different time periods, with in-sample and out-of-sample periods of different lengths.

The results of this analysis appear in Table 3, which begins by analyzing alpha spreads (in terms of basis points per month) in the regression

$$A_{jt} = a_j + b D_{jt}^{OOS} + \epsilon_{jt},$$

where A_{jt} is the alpha spread of predictor j and D_{jt}^{OOS} is an indicator variable that takes the value 1 if t is outside the original sample period used in the paper in which predictor j was first proposed. We include results both with and without predictor fixed effects (a_j). When predictor fixed effects are included, standard errors are clustered by date to account for contemporaneous cross-correlation between predictor alphas. When fixed effects are not included, we additionally cluster by predictor to account for persistent differences across predictors.

First focusing on the results that use predictor fixed effects, the first regression in each panel shows, for Q5-Q1 and CSR, the declines from the in-sample to out-of-sample periods, which are −11.82 and −4.47, respectively. The value −11.82 indicates that the average high minus low spread from quintile sorts declines by almost 12 bps, on a monthly basis, outside the original study's sample period. The value −4.47 means that the impact of a 1-standard-deviation increase in a predictor is smaller by 4.47 basis points per month out of sample. These values are somewhat smaller than those implied by Table 2, but they are nevertheless highly significant.

Next, we separate the out-of-sample period into presample and post-sample subperiods to distinguish data snooping from alternative hypotheses. Following Linnainmaa and Roberts (2018), data snooping should result in lower alpha spreads in both subperiods, while an explanation based on learning would imply a decline only after the sample end. Consistent with the results from Table 2, alpha spreads are lower post-sample but higher presample, though the presample effect is only statistically significant for Q5-Q1.

As in McLean and Pontiff (2016), if investors or fund managers learn from academic studies, then the publication of those studies may lead to a further deterioration in alpha spreads. In Table 3, we investigate publication effects first by replacing the out-of-sample indicator D_{jt}^{OOS} with a post-publication indicator D_{jt}^{PP}, finding post-publication declines that come close to matching the post-sample declines and that are slightly larger than out-of-sample effects. We then include both out-of-sample and post-publication indicators, so that the post-publication indicator captures the incremental decline in alpha following publication, over and above whatever decline occurs out of sample. We find that both coefficients are negative for both Q5-Q1 and CSR, though the post-publication effect is larger and more highly significant, suggesting a greater role for learning relative to data snooping.

Table 3
Predictor panel regressions of out-of-sample effects

A. Q5 - Q1

Intercept							22.26	22.26	21.09	22.26	21.46	22.26
							(6.41)	(6.41)	(5.70)	(6.41)	(5.63)	(6.41)
OOS indicator	−11.82			−6.02		−3.93	−8.34			−3.95		−2.69
	(−3.89)			(−1.94)		(−1.29)	(−2.11)			(−0.64)		(−0.44)
Presample indicator		12.00						9.59				
		(2.15)						(0.75)				
Post-sample indicator		−15.93						−13.00				
		(−4.73)						(−3.98)				
Post-publication indicator			−15.35	−11.19					−11.38	−8.59		
			(−4.44)	(−3.17)					(−3.22)	(−1.35)		
Post-sample decay					−0.10	−0.09					−0.07	−0.06
					(−5.04)	(−4.21)					(−3.56)	(−1.79)
Adj. R^2 (%)	1.13	1.59	1.21	1.25	1.35	1.36	0.17	0.55	0.23	0.24	0.30	0.30
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No

B. CSR

Intercept							7.67	7.67	7.25	7.67	7.39	7.67
							(5.77)	(5.77)	(5.08)	(5.77)	(5.05)	(5.77)
OOS indicator	−4.47			−2.40		−1.35	−2.84			−1.43		−0.96
	(−4.33)			(−2.15)		(−1.25)	(−1.89)			(−0.55)		(−0.38)
Presample indicator		2.86						2.21				
		(1.36)						(0.40)				
Post-sample indicator		−5.79						−4.38				
		(−5.09)						(−3.83)				
Post-publication indicator			−5.76	−4.07					−3.93	−2.91		
			(−4.91)	(−3.24)					(−2.88)	(−1.07)		
Post-sample decay					−0.04	−0.04					−0.03	−0.02
					(−5.82)	(−4.80)					(−3.46)	(−1.52)
Adj. R^2 (%)	1.25	1.55	1.31	1.35	1.48	1.49	0.13	0.37	0.18	0.19	0.24	0.24
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No

This table reports the results of predictor-level panel regressions in which the dependent variable is the alpha spread, in basis points per month, associated with a given predictor. To construct these values, we first compute the realized Carhart 4-factor model alpha of each fund using 36 months of lagged returns to estimate the betas of each fund. We then compute the alpha spread for a given predictor as the difference between the alphas of the highest and lowest quintiles formed based on that predictor (Q5-Q1) or as the slope coefficient of a single cross-sectional regression of fund alphas on the normalized predictor (CSR). The out-of-sample (OOS) or post-publication indicators take the value of 1 if the observation for a given predictor is past the last date in the sample or the publication date of the original paper documenting the significance of that predictor. Post-sample decay is a variable that is equal to $\max\{t-\tau_j,0\}$, where τ_j is the last month in the original sample for predictor j. Regressions with predictor fixed effects are clustered by date when computing standard errors, whereas regressions without fixed effects are clustered by date and predictor. t-statistics are in parentheses. The sample period starts in 1961:11 and ends in 2015:01. There are 8,323 observations for the Q5-Q1 regressions and 8,441 observations for the CSR regressions.

An alternative view of learning is that alpha spreads are likely to decline gradually, perhaps starting to decline just before publication and continuing to decay as awareness of the study grows over time. We attempt to capture this possible dynamic by adding a regressor that captures gradual decay,

$$\max\{t-\tau_j, 0\},$$

where τ_j is the last in-sample period of predictor j. Regressions using this "post-sample decay" variable result in negative coefficients, implying that alpha spreads decrease gradually following the end of the original sample period. Adding the OOS indicator to the regression has little effect, suggesting that the post-sample decay provides a reasonable description of the behavior of alpha spreads outside the sample period.

The table also includes results without predictor fixed effects, which are slightly weaker and show the loss of statistical significance in some cases. Nevertheless, these regressions provide strong evidence of a decline in alpha spreads following the end of the sample period, whether measured using the post-sample or post-publication dummies or the post-sample decay variable.

Overall, the results in Table 3 reinforce the less formal ones in Table 2. Out-of-sample spreads in alphas are much smaller than their in-sample counterparts, with the decline appearing to come entirely from post-sample periods. We find weak evidence of *higher* alpha spreads in presample periods and some evidence of a gradual decline following the end of the original sample period. Because data snooping implies lower alpha spreads both before and after the original sample and does not predict a gradual decline in spreads over time, the evidence from this table is more consistent with learning.

2.3 Fund panel regressions

In this section we analyze out-of-sample predictability of mutual fund alphas in a fund-level panel regression. We do so given that individual funds are likely to display substantial variation in alpha predictors relative to fund portfolios, such as those created as the result of a quintile sort. With larger variation in expected alpha, we may better be able to detect when realized alphas fall short of prior values. In examining individual funds, we will also find cases in which a fund displays multiple characteristics associated with good performance. This will also increase predicted alpha and again give us the best chance to detect any change in the relationship between predicted and realized alphas.

To understand our fund-level regressions, first imagine a regression that we might run if we were using just a single alpha predictor:

$$\alpha_{it} = a + b S_{it} + c S_{it} D_t^{OOS} + d S_{it} D_t^{PP} + \epsilon_{it}.$$

In this regression, S_{it} represents a *score* computed based on fund i's date-t value of the predictor (which is assumed known prior to date t). As an example, the score might take the value -1 if the fund was in the bottom quintile based on the predictor and +1 if it was in the top quintile, with funds in the middle three quintiles receiving a score of 0. As before, D_t^{OOS} and D_t^{PP} are out-of-sample and post-publication indicator variables. Thus, $c=0$ corresponds to the case in which the out-of-sample impact of the predictor is unchanged from its in-sample impact, while $d=0$ indicates the absence of a publication effect.

Note that our regression omits the direct effects of D_t^{OOS} and D_t^{PP}, which are included only when interacted with the score variable. This is because our regressions also include time fixed effects, which will absorb the two indicators and any other variable that exhibits variation only over time. Our results are nearly unchanged when we remove time fixed effects, however, whether or not the direct effects of the two indicators are included.

With multiple predictors, we must make an assumption about how predictability aggregates. We do so by simply assuming that alphas are related to the average score across all predictors. If there are N predictors, we assume that

$$\alpha_{it}=a+b\frac{1}{N}\sum_{j=1}^{N}S_{ijt}+c\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{OOS}+d\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{PP}+\epsilon_{it}, \tag{1}$$

where the j coefficient denotes the predictor. Under this specification, it remains true that $c=0$ implies no deterioration in predictive ability out of sample and that $d=0$ implies no further decline following publication.

We consider two different specifications of the score variable S_{ijt}. One is the extreme quintile score described above, where funds in the bottom predictor quintile receive a score of -1 and funds in the top quintile receive a score of $+1$. This corresponds roughly to the Q5-Q1 approach in our earlier results. The other score is equal to the percentile of the predictor of a given fund, rescaled to lie between -1 and $+1$, within the contemporaneous cross-section. This is more related to, but not analogous to, the earlier CSR results. We have experimented with other definitions of the score and find little effect on our results.

To control for persistent differences between funds, our first specifications include fund fixed effects as well as time fixed effects. As discussed in Pástor, Stambaugh, and Taylor (2015), the use of fund fixed effects in panel regression leads to potential biases, though the Online Appendix shows that these biases are likely to be small in our regressions. Nevertheless, we also report specifications in which we use style-by-time fixed effects in place of time and fund fixed effects.[8] When we do so, we also add controls for fund fees (bps per month) and size (log of AUM), which are common in the literature, to our regressions, and we remove these two variables from the list of predictors used to calculate fund scores (the $S_{i,j,t}$ from equation (1)). To a large extent, fund fixed effects appear to proxy for expense ratios, which are very persistent, and may also proxy for

[8] The Online Appendix includes a number of robustness checks for this table and others. These checks include variation in the specification of fixed effects.

scale effects. Thus, these two controls are potentially adequate replacements for fund fixed effects.[9]

While these alternative specifications alleviate concerns about bias, their interpretation is somewhat different from regressions that exclude controls for fund fee and size. Without size and fee controls, the post-publication coefficient, for instance, measures the effect of publication on alphas net of fees and costs. With those controls, the same coefficient is more interpretable as the effect of publication on managerial ability, in the sense of Berk and Green (2004), in that it has been cleansed of scale and fee effects. While our primary focus is on the dynamics of net alphas, we believe it is also interesting to study how the part of fund alphas unrelated to fees and size evolves over time.

The results, shown in Table 4, are generally consistent with those in the predictor panel. With fund and time fixed effects, out-of-sample effects are significant and large in terms of economic magnitude. For example, in the first column of panel A, the -64.29 coefficient on Average Score $\times$ OOS Indicator (the c coefficient from (1)) implies that a fund with a predictor average of 0.2 (which is approximately at the 75th percentile) will see its monthly alpha fall by $64.29 \times 0.2 \approx 13$ bps, or 1.5% annualized, during the out-of-sample period. As with the predictor panel, when we separate the presample and post-sample periods, it is only the latter where we find statistically significant evidence of lower predictability. Similarly, the post-publication effect and post-sample decay effect are both significant when these variables are included one at a time.

A noticeable difference from the predictor panel results in Table 3 is the inability to separate publication effects from out-of-sample effects, which following McLean and Pontiff (2016) is a way to distinguish between data mining and learning hypotheses. With fund and time fixed effects, we find no significant presample effect. Results are mixed when the out-of-sample indicator is included alongside the post-publication indicator or post-sample decay.

Specifications using style-by-time fixed effects are weaker, suggesting that out-of-sample and post-publication changes in fund alphas are primarily a within-fund phenomenon and that persistent fund-level differences are not perfectly captured by our controls. Nevertheless, we still find consistent results with these specifications. In particular, the post-sample reduction in alphas is statistically significant in all specifications, though only marginally so in one. Coefficient magnitudes are also lower in these specifications, but they are typically at least half the size of those from the regressions with fund fixed effects and are economically significant.

To summarize the evidence presented in this section, we find strong support for lower post-sample alpha predictability using a variety of methodologies, but

9 As discussed in Pástor, Stambaugh, and Taylor (2015), the endogeneity of fund size prevents a clear interpretation of its coefficient; however, this is not our goal.

Table 4
Fund panel regressions of out-of-sample effects

A. Percentile scores

Average score (*b*)	49.32 (3.56)	45.58 (3.20)	34.10 (3.81)	48.51 (3.42)	33.25 (3.37)	46.46 (3.30)	52.40 (3.25)	52.23 (3.12)	39.36 (3.69)	52.82 (3.21)	38.42 (3.38)	53.03 (3.29)
Average score × OOS indicator (*c*)	−64.29 (−4.01)			−45.89 (−1.75)		−38.33 (−1.74)	−38.73 (−1.93)			−46.98 (−1.53)		−43.72 (−1.66)
Average score × Presample indicator (*d*)		23.19 (0.84)						−33.86 (−0.66)				
Average score × Post-sample indicator (*e*)		−63.17 (−3.93)						−38.58 (−1.89)				
Average score × Post-pub indicator (*f*)			−63.01 (−4.68)	−26.44 (−1.15)					−28.71 (−1.70)	11.26 (0.46)		
Average score × Post-sample decay (*g*)					−0.34 (−4.76)	−0.19 (−2.00)					−0.15 (−1.52)	0.03 (0.27)
Expense ratio (*h*)							−0.93 (−3.30)	−0.93 (−3.32)	−0.88 (−3.13)	−0.93 (−3.31)	−0.89 (−3.19)	−0.93 (−3.31)
log TNA (*i*)							−0.87 (−2.23)	−0.87 (−2.23)	−0.88 (−2.18)	−0.90 (−2.24)	−0.84 (−2.00)	−0.89 (−2.11)
Adj. R^2 (%)	10.48	10.49	10.46	10.48	10.47	10.48	13.87	13.87	13.85	13.87	13.85	13.87
Fixed effects	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Style × time	Style × time	Style × time	Style × time	Style × time	Style × time

(Continued)

Table 4
(Continued)

B. Extreme quintile scores

Average score (b)	41.48	39.72	28.10	41.21	28.76	39.74	46.64	46.45	34.63	46.85	34.55	46.93
	(3.64)	(3.45)	(3.72)	(3.59)	(3.41)	(3.48)	(3.48)	(3.43)	(3.93)	(3.47)	(3.58)	(3.55)
Average score × OOS indicator (c)	−53.06			−39.46		−31.21	−33.08			−39.20		−35.95
	(−4.12)			(−2.11)		(−1.97)	(−1.99)			(−1.76)		(−1.91)
Average score × Presample indicator (d)		11.73						−25.17				
		(0.61)						(−0.84)				
Average score × Post-sample indicator (e)		−54.41						−32.97				
		(−4.22)						(−1.98)				
Average score × Post-pub indicator (f)			−50.19	−20.53					−22.69	8.68		
			(−4.63)	(−1.29)					(−1.66)	(0.53)		
Average score × Post-sample decay (g)					−0.29	−0.17					−0.13	0.02
					(−4.81)	(−2.43)					(−1.49)	(0.23)
Expense ratio (h)							−0.82	−0.81	−0.77	−0.82	−0.77	−0.82
							(−2.82)	(−2.82)	(−2.61)	(−2.82)	(−2.64)	(−2.79)
log TNA (i)							−0.79	−0.78	−0.81	−0.81	−0.77	−0.80
							(−2.02)	(−2.02)	(−2.04)	(−2.05)	(−1.83)	(−1.91)
Adj. R^2 (%)	10.47	10.48	10.45	10.47	10.46	10.47	13.86	13.86	13.84	13.86	13.85	13.86
Fixed effects	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Style × time	Style × time	Style × time	Style × time	Style × time	Style × time

This table reports the results of fund-level panel regressions that analyze the out-of-sample and post-publication performance of mutual fund predictors. All regressions are restricted versions of

$$\alpha_{it}=a+b\frac{1}{N}\sum_{j=1}^{N}S_{ijt}+c\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{OOS}+d\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{Pre}+e\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{Post}+f\frac{1}{N}\sum_{j=1}^{N}S_{ijt}D_{jt}^{PP}+g\frac{1}{N}\sum_{j=1}^{N}S_{ijt}\max\{t-\tau_j,0\}+h\text{Fee}_{it}+i\log\text{TNA}_{it}+\epsilon_{it},$$

where α_{it} is the realized alpha in basis points for fund i in month t; S_{ijt} is fund i's score on predictor j in month t; and the D_{jt} variables are indicators for whether predictor j is out-of-sample (OOS), presample (Pre), post-sample (Post), or post-publication (PP) in month t. τ_j is the last month of the in-sample period for predictor j, and N is the number of predictors considered. Alphas are inclusive of all fund fees and costs. In panel A, scores are equal to the percentile of each fund's predictors relative to the contemporaneous cross-section, rescaled to [−1,+1]. In panel B, scores are equal to +1 for predictors in the top quintile and −1 for predictors in the bottom quintile. For regressions with fund and time fixed effects, we cluster by date when computing standard errors. The sample for these regressions is from 1961:11 to 2015:01 and includes 356,873 observations. For regressions with style-by-time fixed effects, we cluster by fund and date, include controls for expense ratio (Fee) and log TNA, and exclude those two variables from the calculation of fund scores. These regressions use a sample starting in 1980:01 that includes 265,860 observations. t-statistics are in parentheses.

neither the data snooping nor the learning hypotheses appear fully consistent with the decline. Data snooping should imply lower predictability both before and after the original samples, but we find no presample decline and some evidence of a presample increase in predictive ability. Learning appears to receive slightly greater support based on its ability to explain the one-sided decline in predictability. However, additional results in the Online Appendix show that the predictions of the learning hypothesis for fund fees and asset growth are not upheld in the data. Together, these findings suggest that out-of-sample effects may have a different explanation.

3. Time-Varying Arbitrage Activity and Fund Competition

A competing explanation for diminished out-of-sample alpha spreads is that rising market efficiency has made alpha generation (either positive or negative) more difficult over time. This is the conclusion of Chordia, Subrahmanyam, and Tong (2014), who show that measures of the intensity of arbitrage activity are inversely correlated with the profitability of many of the best-known asset pricing anomalies. Of the 12 anomalies they consider, 11 are decreasing in the amount of hedge fund assets under management, 10 are decreasing in the level of aggregate short interest, and 11 are decreasing in aggregate share turnover. In this section we follow Chordia, Subrahmanyam, and Tong by examining the relation between arbitrage activity and market efficiency, but where we measure the latter based on spreads in mutual fund alphas.

Alternatively, the ability of funds to generate nonzero alpha may have declined due to an increase in the degree of competition between mutual funds, a hypothesis first considered by Pástor and Stambaugh (2012). Early in our sample, active U.S. mutual funds collectively held less than 1% of the aggregate market capitalization of CRSP stocks, while by 2015 that value had reached 15%. If fund returns are diminishing with the scale of fund industry, as Pástor, Stambaugh, and Taylor (2015) conclude, then perhaps the increased competition that results from greater scale also reduces the ability of the alpha predictors we analyze. There is also likely cross-sectional variation in competition. Wahal and Wang (2011) show that the performance of incumbent mutual funds is reduced by the entrance of new funds holding similar portfolios. Hoberg, Kumar, and Prabhala (2018) find that funds in more competitive segments of style space are less likely to outperform. Competition-related compression in alphas could also decrease the effectiveness of the alpha predictors we consider.

3.1 Arbitrage and alpha

Following Chordia, Subrahmanyam, and Tong (2014), we consider three different proxies of market-level arbitrage activity – aggregate short interest, aggregate share turnover, and aggregate hedge fund asset size. Aggregate short interest of U.S. common equities is the value-weighted monthly short interest scaled by the previous month's outstanding shares. Stock short interest data is from Compustat, and our final aggregate short interest spans from January 1973 to December 2014. Aggregate share turnover is the monthly value-weighted share turnover using the market capitalization at the end of the previous year as the weight. Monthly share turnover is share trading volume scaled by shares outstanding, which are all on U.S. common equities and are obtained from CRSP. Our final aggregate share turnover spans from January 1961 to December 2014. Aggregate hedge fund asset size is hedge fund monthly assets under management (AUM) scaled by the market capitalization of NYSE and AMEX stocks in the previous month. Hedge fund monthly AUM from March 1977 to December 2010 are from Jylhä and Suominen (2011). We extend this series through 2014 using annual AUM data from Hedge Fund Research. Our final aggregate hedge fund asset size therefore spans from March 1977 to December 2014 (allowing alpha forecasts through January 2015).

In Chordia, Subrahmanyam, and Tong (2014), each of the arbitrage activity proxies is detrended by regressing the measure on a time trend prior to its inclusion in a regression. We believe that detrended measures are useful in that they potentially identify effects that cannot be explained by a simple trend toward market efficiency. At the same time, if markets have become more efficient over time as the result of an overall positive trend in arbitrage activity, then detrending will remove some of the variation driving alpha dynamics. While economic arguments would appear to favor not detrending, we consider both detrended and nondetrended series.

In addition, we create a composite arbitrage activity series by averaging across the three measures. Because short interest and hedge fund AUM are only available starting in 1973 and 1977, respectively, the composite measure only includes turnover before 1973 and only includes turnover and short interest before 1977. Prior to averaging, we divide each arbitrage measure by its own standard deviation, so that each series contributes about the same to the aggregate measure. We also shift the means of the short interest and hedge fund AUM series so that the initial value of each series is zero, which eliminates discontinuities in the composite measure when each of those component series is introduced. One average is computed for detrended arbitrage proxies, and one average is computed for nondetrended proxies.

We investigate the role of arbitrage activity within the fund panel framework described in Section 2.3. In this approach, we compute the average score,

$$\mathrm{Avg}S_{it} = \frac{1}{N}\sum_{j=1}^{N} S_{ijt},$$

for each fund i at each time t and run the regression

$$\alpha_{it} = a + b\,\mathrm{Avg}S_{it} + c\,Z_t\,\mathrm{Avg}S_{it} + \epsilon_{it},$$

where Z_t, known at the start of date t, is one of the arbitrage activity proxies or composite measures discussed above.[10] A negative c coefficient indicates that greater arbitrage activity reduces the strength of the predictive relationship between average scores and future returns. We report results with time and fund fixed effects and with style-by-time fixed effects.

The results, in Table 5, show that arbitrage effects are significant in every regression. For all proxies, both with and without detrending, higher arbitrage activity is associated with a reduction in predictability, in that predictors are significantly less positively related to fund alphas when arbitrage activity is high. This result holds for both percentile scores and scores based on extreme quintiles and for both specifications of fixed effects. Untabulated results based on time fixed effects only are also extremely similar.

Furthermore, the coefficient estimates imply economically important effects. Across all specifications, c coefficients range from about -15 to -30. A fund in the 75th percentile—good, but not great—has an average percentile-based score of about 0.2. For that fund, a c coefficient of -20 implies that a 1-standard-deviation increase in arbitrage activity is associated with a 0.5% decrease in annual alphas.[11]

A natural extension is to measure cross-sectional variation in arbitrage intensity and to determine whether that, too, is negatively correlated with alpha predictability. Given the possibility that there exists meaningful variation across fund styles, we compute style-level arbitrage proxies and compare their ability to explain changing alpha spreads with the aggregate measures analyzed previously. While data limitations prevent a disaggregation of our hedge fund AUM measure, it is straightforward to compute measures of short selling and turnover at the style level. When we do so, we find that these measures are highly collinear with the aggregate versions and that it is difficult to detect any marginal effect of style-level arbitrage, either in the predictor panel or fund panel.

Because of their similarity to the results presented here and the results that follow, we do not show the analogous regressions based on the predictor panel approach proposed by McLean and Pontiff (2016) and used in Section 2.2. Those results, which may be found in the Online Appendix, are similarly strong, indicating that arbitrage activity is negatively linked to alpha spreads computed either using quintile sorts or cross-sectional regressions.

3.2 Competition and alpha

The arbitrage proxies that we consider largely measure forces acting from outside the mutual fund industry. However, mutual fund alpha generation may also be affected by the degree of competition within that industry.

[10] We do not estimate the direct effect of Z_t, because it is absorbed by time fixed effects.

[11] The change in alpha following a 1-standard-deviation increase in Z_t is $c \times \text{Avg}S_{it} = -20 \times 0.2 = -4$ bps per month, or a decrease of 48 bps per year.

Table 5
Arbitrage activity

	Fund & time fixed effects			Style × time fixed effects		
Arbitrage activity proxy (Z)	Avg. score (Avg$S$)	$Z \times$ Avg$S$	Adj. R^2 (%) & # of obs.	Avg. score (Avg$S$)	$Z \times$ Avg$S$	Adj. R^2 (%) & # of obs.
A. Percentile scores						
Short interest	67.03	−30.62	10.46	85.41	−27.49	13.92
	(4.11)	(−4.27)	353,504	(4.36)	(−3.24)	265,860
Short interest - detrended	−35.85	−18.68	10.45	−10.11	−16.77	13.93
	(−2.35)	(−2.90)	353,504	(−0.67)	(−2.60)	265,860
Turnover	52.54	−26.60	10.52	69.43	−22.90	13.95
	(4.76)	(−4.39)	356,873	(5.48)	(−3.46)	265,860
Turnover - detrended	−6.11	−17.11	10.49	15.48	−15.77	13.94
	(−0.67)	(−3.44)	356,873	(1.92)	(−3.12)	265,860
Hedge fund AUM	44.61	−24.49	10.17	59.70	−19.31	13.91
	(3.14)	(−3.83)	349,751	(4.41)	(−3.08)	265,860
Hedge fund AUM - detrended	−36.87	−20.50	10.17	−6.78	−16.85	13.91
	(−2.89)	(−3.16)	349,751	(−0.57)	(−2.81)	265,860
Avg. of nondetrended arb. proxies	62.45	−31.69	10.52	77.34	−26.56	13.94
	(4.76)	(−4.87)	356,873	(4.96)	(−3.56)	265,860
Avg. of detrended arb. proxies	−35.76	−25.95	10.51	−12.79	−24.36	13.97
	(−2.67)	(−3.59)	356,873	(−0.97)	(−3.38)	265,860
B. Extreme quintile scores						
Short interest	59.22	−26.90	10.45	77.86	−24.81	13.91
	(4.19)	(−4.30)	353,504	(4.40)	(−3.21)	265,860
Short interest - detrended	−29.96	−15.94	10.44	−7.00	−14.50	13.90
	(−2.20)	(−2.79)	353,504	(−0.50)	(−2.46)	265,860
Turnover	46.34	−23.27	10.51	62.49	−20.20	13.92
	(4.76)	(−4.29)	356,873	(5.33)	(−3.24)	265,860
Turnover - detrended	−4.52	−14.72	10.48	15.05	−13.66	13.92
	(−0.56)	(−3.26)	356,873	(2.02)	(−2.86)	265,860
Hedge fund AUM	40.14	−21.93	10.16	54.82	−17.55	13.89
	(3.26)	(−3.94)	349,751	(4.52)	(−3.10)	265,860
Hedge fund AUM - detrended	−32.60	−18.26	10.16	−5.16	−15.06	13.90
	(−2.91)	(−3.23)	349,751	(−0.48)	(−2.81)	265,860
Avg. of nondetrended arb. proxies	55.32	−27.95	10.51	70.19	−23.80	13.92
	(4.89)	(−4.92)	356,873	(4.99)	(−3.49)	265,860
Avg. of detrended arb. proxies	−29.89	−22.26	10.49	−9.74	−21.27	13.94
	(−2.51)	(−3.48)	356,873	(−0.80)	(−3.23)	265,860

This table reports the results of fund-level panel regressions that analyze the relation between realized mutual fund alphas and various proxies for the level of arbitrage activity. The regression equation is

$$\alpha_{it} = a + b\,\text{Avg}S_{it} + c\,Z_t\,\text{Avg}S_{it} + \epsilon_{it},$$

where α_{it} is the realized alpha in basis points for fund i in month t, Avg S_{it} is fund i's average percentile-based or extreme quintile-based score across all predictors in month t, and Z_t is an arbitrage activity proxy observable at the start of month t. Arbitrage proxies include the aggregate short interest of U.S. common equities, the stock market's turnover ratio, and the ratio of total assets under management of U.S. hedge funds to the U.S. market capitalization. These proxies are included in raw and detrended forms. The regression also includes equal weighted averages of the three arbitrage proxies. For regressions with fund and time fixed effects, we cluster by date when computing standard errors. For regressions with style-by-time fixed effects, we cluster by fund and date, include controls for expense ratio and log TNA (coefficients unreported), and exclude those two variables from the calculation of fund scores. t-statistics are in parentheses. The sample start date is 1961:11 for regressions based on turnover and average proxies, 1973:01 for short interest, and 1977:03 for hedge fund AUM. All samples end in 2015:01.

Several recent studies provide support for this hypothesis. Following Pástor and Stambaugh (2012), Pástor, Stambaugh, and Taylor (2015) show strong evidence that mutual fund performance is decreasing in the aggregate amount of assets under management of all actively managed U.S. equity mutual funds. They also analyze the effects of changing AUM at the style level, aggregating all funds within a given size/value category, but find no evidence that this measure is related to fund performance. In contrast, Hoberg, Kumar, and Prabhala (2018) find evidence of style-related competition by analyzing the density of competitors in the specific size/value/momentum region in which a given fund is active. Their approach avoids arbitrary style cutoffs and accounts for both the number and similarity of competing funds.

Following these papers, we consider four different measures of competition and investigate how they affect alpha predictability. The first is aggregate industry size, defined as in Pástor, Stambaugh, and Taylor (2015) as the total AUM of all actively managed U.S. equity mutual funds divided by the total market capitalization of all CRSP common stocks. The second is sector size, computed similarly but separately for each style. As described earlier, styles are determined based on two-by-two size/value grid, where size and value are measured by the holdings-weighted average percentiles of the stocks in each fund. The third and fourth are the *NPeers* (number of peers) and *TSIM* (total similarity) measures of Hoberg, Kumar, and Prabhala (2018). The former measures only the number of peers with holdings of similar size/value/momentum, while the latter takes into account the degree of similarity. Note that these measures, with the exception of industry size, are different from our arbitrage proxies in that they vary cross-sectionally as well as through time.

The results of this analysis, in Table 6, show that all four competition measures have a negative and significant relation with alpha predictability when included individually. Regressions that include all four variables appear to suffer from some redundancy among the regressors, but bivariate specifications seem to favor roles for industry size and either of the two competition measures of Hoberg, Kumar, and Prabhala (2018).[12]

To understand the magnitude of these effects, Pástor, Stambaugh, and Taylor (2015) note that a 0.01 increase in industry size occurs every few years, on average. Because a fund in the 75th percentile has an average percentile-based score of about 0.2, using the industry size coefficient estimate of -301.23 from the first regression in panel A, we can estimate that a 0.01 increase in industry size reduces that fund's monthly alpha by approximately 0.6 bps ($-301.23 \times 0.01 \times 0.2$). In comparison, Pástor, Stambaugh, and Taylor (2015) estimate that

12 When computing NPeers and TSIM, we impose the same data requirements used by Hoberg, Kumar, and Prabhala (2018). The result is that not every fund in our main sample has a value for these variables. This explains the discrepancy in the number of observations used in different regressions in Table 6.

Table 6
Competition

A. Percentile scores

Avg S	68.76	55.52	38.56	36.79	66.58	67.77	67.44	67.21
	(5.34)	(6.12)	(4.75)	(4.75)	(5.31)	(5.28)	(5.38)	(5.36)
Avg S × ISize	−301.23				−177.91	−180.93	−229.05	−235.35
	(−3.13)				(−1.58)	(−1.59)	(−2.25)	(−2.34)
Avg S × SSize		−265.84			−74.69	−145.13		
		(−3.27)			(−0.80)	(−1.56)		
Avg S × NPeers			−0.12		0.03		−0.09	
			(−3.91)		(0.19)		(−2.78)	
Avg S × TSIM				−0.97	−0.94			−0.73
				(−4.03)	(−0.80)			(−2.96)
Adj. R^2 (%)	13.87	13.87	14.17	14.17	14.19	13.88	14.19	14.19
# of obs.	265,860	265,860	233,410	233,410	233,410	265,860	233,410	233,410

B. Extreme quintile scores

Avg S	64.60	50.91	34.24	32.90	62.69	63.71	63.49	63.31
	(5.38)	(6.13)	(4.63)	(4.67)	(5.40)	(5.33)	(5.47)	(5.45)
Avg S × ISize	−284.13				−195.90	−186.19	−230.60	−234.07
	(−3.18)				(−1.88)	(−1.76)	(−2.47)	(−2.54)
Avg S × SSize		−240.48			−56.40	−117.67		
		(−3.20)			(−0.64)	(−1.36)		
Avg S × NPeers			−0.11		0.08		−0.07	
			(−3.56)		(0.53)		(−2.37)	
Avg S × TSIM				−0.83	−1.20			−0.59
				(−3.71)	(−1.08)			(−2.58)
Adj. R^2 (%)	13.86	13.86	14.16	14.16	14.18	13.87	14.18	14.18
# of obs.	265,860	265,860	233,410	233,410	233,410	265,860	233,410	233,410

This table reports the results of fund-level panel regressions that analyze the relation between realized mutual fund alphas and various proxies for the level of competition faced by the mutual fund. The unrestricted regression equation is

$$\alpha_{it} = a + b\,\mathrm{Avg}S_{it} + c\,\mathrm{ISize}_t\,\mathrm{Avg}S_{it} + d\,\mathrm{SSize}_t\,\mathrm{Avg}S_{it} + e\,\mathrm{NPeers}_{i,t}\,\mathrm{Avg}S_{it} + f\,\mathrm{TSIM}_{i,t}\,\mathrm{Avg}S_{it} + \epsilon_{it},$$

where α_{it} is the realized alpha in basis points for fund i in month t, and Avg S_{it} is fund i's average percentile-based or extreme quintile-based score across all predictors in month t. The first two competition proxies, from Pástor, Stambaugh, and Taylor (2015), are the size of the mutual fund industry (ISize) and the size of the mutual fund's sector (SSize), both as a fraction of aggregate/sector market capitalization. The other two proxies, from Hoberg, Kumar, and Prabhala (2018), are the number of peers of the fund (NPeers) and the total similarity among peers (TSIM). All regressions include style-by-time fixed effects. We also include controls for expense ratio and log TNA (coefficients unreported), and those variables are excluded from the calculation of fund scores. t-statistics, which are in parentheses, are clustered by date and fund. The sample period starts in 1980:1 and ends in 2015:01.

the same increase in industry size would reduce the average fund alpha by about 3 bps per month.

As in Hoberg, Kumar, and Prabhala (2018), a fund in the high competition tercile may have 100 more peers than one in the low competition tercile. For a fund with an average percentile-based score of 0.2, the third regression in panel A implies that the effects of such a change in the number of peers would be approximately 2.4 bps ($-0.12 \times 100 \times 0.2$) per month. In contrast, Hoberg, Kumar, and Prabhala (2018) estimate that the difference in average alphas between the high and low terciles is about 7–8 bps per month. Thus, competitive effects on alpha spreads are significant, but their magnitude is modest relative to the estimated effects on average alphas.

For brevity, Table 6 only included results based on style-by-time fixed effects. Results based on time fixed effects only are extremely similar. Adding fund fixed effects naturally reduces the significance of the Hoberg, Kumar, and Prabhala (2018) competition measures. The importance of industry size is also heightened under fund fixed effects. Overall, however, the additional results paint a consistent picture relative to those presented here.

3.3 The post-sample effect revisited

In this section we ask whether it is possible that the declining post-sample predictability is consistent with time-varying arbitrage activity or mutual fund competition, rather than data snooping or learning effects. To answer this question, we consider specifications that include both a post-sample effect and measures of arbitrage activity and/or competition. If the post-sample effect remains significant, this would indicate that the sample period of the original study is important, either due to data mining or learning. If including arbitrage or competition measures drives out post-sample effects, then we may conclude that declining predictability in fund alphas has less to do with data snooping or learning than it does with an overall rise in market efficiency or competition.

We consider these issues within the fund panel setting by estimating full and restricted versions of the specification

$$\alpha_{it} = a + b\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + c\frac{1}{N}\sum_{j=1}^{N} S_{ijt}\max\{t-\tau_j, 0\} + d\,Z_t\frac{1}{N}\sum_{j=1}^{N} S_{ijt}$$

$$+ e\,\mathrm{ISize}_t\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + f\,\mathrm{TSIM}_{i,t}\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + g\,t\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + \epsilon_{it}, \qquad (2)$$

where the b coefficient captures the baseline relation between average predictor scores and alpha, c measures post-sample decay, d captures the relation with our arbitrage activity measure Z_t, e and f assess the effects of industry size (ISize) and the total similarity (TSIM) measure of Hoberg, Kumar, and Prabhala (2018), and g captures the relation between average scores and a time trend. We include regressions with time and fund fixed effects and with style-by-time fixed effects.

Several consistent results emerge from the regression results, which are reported in Table 7. The first is that all evidence for post-sample decay vanishes upon controlling for arbitrage activity, which remains significant in all regressions, with coefficient estimates that are similar or larger than those in Table 5. Industry size explains some fraction of the post-sample variation in realized alphas, but it is subsumed by arbitrage activity when both variables are included. Finally, the significance of TSIM is unaffected by the inclusion of other predictors, though only under specifications that do not include fund fixed effects.

Table 7
Comparison of alternative hypotheses

A. Percentile scores

Average score (b)	33.25	62.02	45.94	37.65	−83.00	38.42	80.67	65.80	55.52	−16.92
	(3.37)	(4.64)	(3.20)	(2.89)	(−1.19)	(3.38)	(5.19)	(5.24)	(4.99)	(−0.26)
Average score × Post-sample decay (c)	−0.34	−0.03	−0.26	−0.02	−0.19	−0.15	0.20	−0.03	0.18	0.10
	(−4.76)	(−0.26)	(−2.02)	(−0.19)	(−1.28)	(−1.52)	(1.43)	(−0.24)	(1.24)	(0.56)
Average score × Arbitrage (d)		−30.19		−49.17	−58.64		−36.75		−49.03	−55.99
		(−3.06)		(−2.70)	(−3.04)		(−3.49)		(−3.02)	(−3.12)
Average score × ISize (e)			−138.24	449.90	31.36			−205.45	414.09	158.98
			(−0.78)	(1.64)	(0.08)			(−1.29)	(1.68)	(0.48)
Average score × TSIM (f)			−0.59	−0.57	−0.57			−0.73	−0.71	−0.71
			(−1.89)	(−1.81)	(−1.81)			(−2.93)	(−2.83)	(−2.85)
Adj. R^2 (%)	10.47	10.52	10.18	10.25	10.27	13.85	13.95	14.19	14.28	14.29
# of obs.	356,873	356,873	274,924	274,924	274,924	265,860	265,860	233,410	233,410	233,410
Fixed effects	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Style × time	Style × time	Style × time	Style × time	Style × time
Controls					trend	fee, size	fee, size	fee, size	fee, size	fee, size, trend

(Continued)

Table 7
(Continued)

B. Extreme quintile scores

Average score (b)	26.94	54.67	44.95	36.01	−56.32	34.55	72.72	62.41	52.55	−9.41
	(3.41)	(4.78)	(3.46)	(3.04)	(−0.96)	(3.58)	(5.20)	(5.46)	(5.14)	(−0.16)
Average score × Post-sample decay (c)	−0.29	−0.04	−0.20	−0.03	−0.14	−0.13	0.15	−0.02	0.13	0.07
	(−4.81)	(−0.54)	(−2.00)	(−0.35)	(−1.32)	(−1.49)	(1.33)	(−0.19)	(1.17)	(0.54)
Average score × Arbitrage (d)		−25.66		−41.29	−49.78		−31.12		−40.08	−46.54
		(−3.25)		(−2.66)	(−2.92)		(−3.55)		(−2.83)	(−2.88)
Average score × ISize (e)			−174.64	347.92	22.59			−215.47	312.56	92.99
			(−1.18)	(1.42)	(0.07)			(−1.63)	(1.42)	(0.32)
Average score × TSIM (f)			−0.49	−0.45	−0.46			−0.60	−0.56	−0.57
			(−1.72)	(−1.61)	(−1.63)			(−2.55)	(−2.40)	(−2.44)
Adj. R^2 (%)	9.77	9.82	10.17	10.23	10.24	13.85	13.93	14.18	14.25	14.25
# of obs.	356,873	356,873	274,924	274,924	274,924	265,860	265,860	233,410	233,410	233,410
Fixed effects	Fund & time	Fund & time	Fund & time	Fund & time	Fund & time	Style × time	Style × time	Style × time	Style × time	Style × time
Controls					trend	fee, size	fee, size	fee, size	fee, size	fee, size, trend

This table reports the results of fund-level panel regressions of full and restricted versions of

$$\alpha_{it} = a + b\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + c\frac{1}{N}\sum_{j=1}^{N} S_{ijt}\max\{t-\tau_j,0\} + d\,Z_t\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + e\,\text{ISize}_t\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + f\,\text{TSIM}_{i,t}\frac{1}{N}\sum_{j=1}^{N} S_{ijt} + \text{controls} + \epsilon_{it},$$

where α_{it} is the realized alpha in basis points for fund i in month t; S_{ijt} is fund i's score on predictor j in month t; τ_j is the last date of the in-sample period for predictor j; Z_t is a composite measure of arbitrage activity; ISize is the total mutual fund industry AUM divided by the aggregate stock market capitalization; and TSIM is the total similarity measure of Hoberg, Kumar, and Prabhala (2018). For regressions with fund and time fixed effects, we cluster by date when computing standard errors and use a sample from 1961:11 to 2015:01. For regressions with style-by-time fixed effects, we cluster by fund and date, include controls for expense ratio and log TNA, exclude those two variables from the calculation of fund scores, and use a sample starting in 1980:01. t-statistics are in parentheses.

These findings are robust. Results from the predictor panel setting are essentially the same as those reported here. Similar conclusions are obtained under a number of alternative specifications for both predictor and fund panels.

Overall, the evidence shows strong support for the influence of time-varying arbitrage activity, though competition appears to be a secondary explanation for changing alpha predictability. No support is found for any post-sample effect, suggesting that any apparent effect is an artifact of model misspecification. Instead, arbitrage forces appear to have strengthened over time, causing fund alphas to gradually shrink toward zero. Regressing alpha spreads on our post-sample decay variable or a post-sample or post-publication indicator results in a negative coefficient, but this is only because of omitted variable bias due to the exclusion of arbitrage effects and/or competition effects. We therefore conclude that there is no convincing evidence that any of the post-sample decline in alpha predictability is due to either data mining or learning.

3.4 Implications

Given that alpha predictability has declined, what does this mean for a mutual fund investor? If the decline were mainly due to a reduction in the number and magnitude of negative alphas, it is not clear that an informed investor would be affected by the patterns we have documented. The relevant question for that investor is whether positive alphas have attenuated or become rarer.

We examine this issue by analyzing the performance of a portfolio of funds that are predicted to have the best future performance. These predictions are based on a model estimated on the full panel of mutual funds. At the beginning of each month, we use lagged information to compute predicted alphas for each fund based on the most general specification estimated (with both fund and time fixed effects) from panel A of Table 7. After sorting on predicted alphas, we form a portfolio of the top quintile of funds. We then calculate the realized alpha of that portfolio by averaging the realized alphas of the constituent funds.

Our interest is in whether these realized alphas have declined over time. Because they are extremely noisy on a month-by-month basis, we employ two different smoothing strategies. One is to examine 10-year moving averages. The other is to examine sample splits, where we measure average alphas before and after various breakpoints.

Figure 1 reveals the results of this analysis. The top panel shows that the top quintile portfolio started our sample period with a highly significant alpha of roughly 30 bps per month. By the end of the sample period that value had fallen to 10 bps per month, with only marginal significance.

The middle panel of the figure shows the results of splitting the sample at various breakpoints. For each breakpoint, whose value appears on the horizontal axis, we compute a "backward" average consisting of all data prior to the breakpoint and a "forward" average consisting of data on or after the breakpoint. Breakpoints are chosen such that each subsample contains at least 12 months.

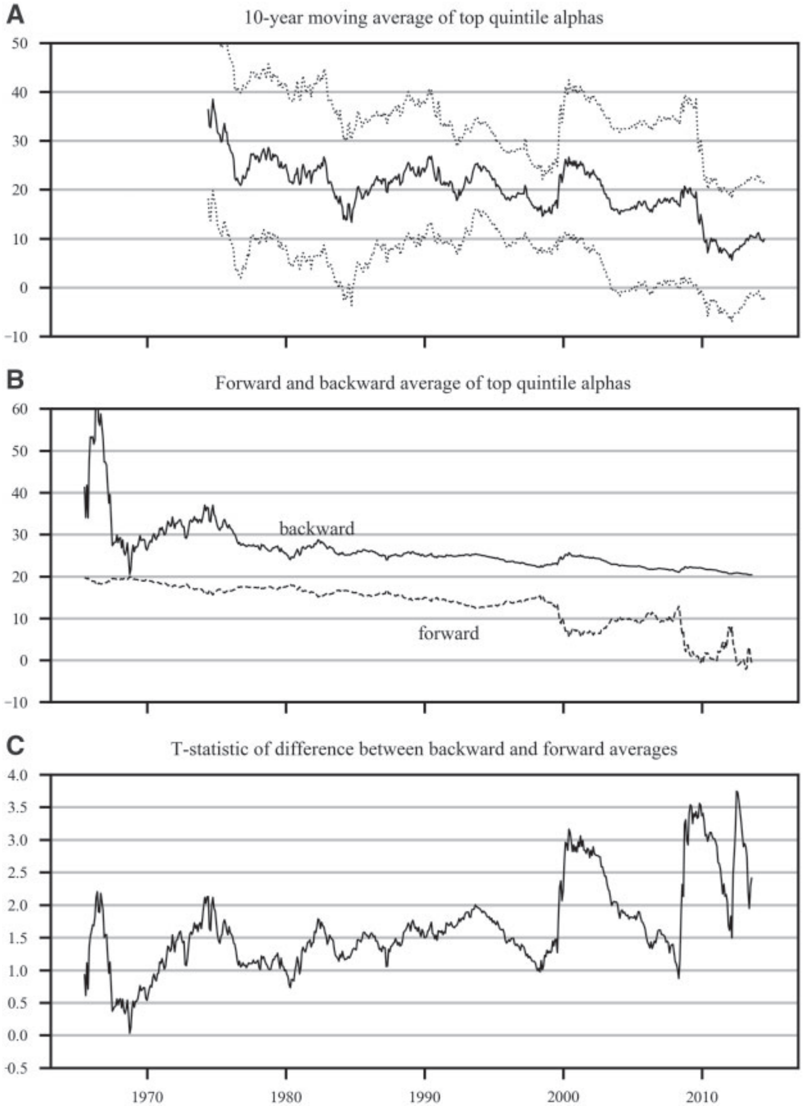

Figure 1
Performance of the top quintile fund portfolio
In each month, we sort mutual funds based on the predicted alphas from the final regression, with both fund and time fixed effects, in panel A of Table 7. We then form a portfolio consisting of all funds in the top quintile. Panel A shows the 10-year moving average of the realized alpha of that portfolio and asymptotic 95% confidence intervals. In panel B, we split the full sample at various breakpoints, whose values appear on the horizontal axis. For each breakpoint, we compute a "backward" average consisting of all portfolio alphas before the breakpoint and a "forward" average of alphas after the breakpoint. Panel C reports t-statistics of the difference between backward and forward averages.

The middle panel shows that backward averages are always greater than forward averages, indicating that regardless of the breakpoint chosen the future performance of the portfolio is always inferior to the past. The last five years of our sample actually display zero realized alpha on average. The bottom panel shows the corresponding t-statistics of the difference between backward and forward averages. For most post-2000 breakpoints, the difference is significant.

These results are robust to a number of possible specification changes. A model that only includes arbitrage effects generates very similar results, consistent with its relative importance in the full model. Results using models based on extreme quintile scores are also very similar.

We conclude that the evidence in favor of positive investment in actively managed equity mutual funds is currently weak, at least for those investors basing decisions on the predictors we have analyzed. While some alpha may remain, it is far below historical averages and may have disappeared completely. Considering that even a well-diversified portfolio of actively managed funds adds at least several percentage points of idiosyncratic risk to a comparable passive portfolio, the risk-return tradeoff inherent in holding the high quintile portfolio appears marginal at best.

4. Corporate Bond Funds

The vast majority of the mutual fund predictors we consider in this paper were originally studied in the context of U.S. equity funds.[13] Were their in-sample significance the result of extensive data snooping, there would be little reason to think that their validity would be maintained in funds holding other securities. To the extent that characteristics from the equity fund universe are predictive in other markets, it is less likely that their original effects were spurious. Examining a different market also allows us to provide corroborating evidence on the importance of time-varying arbitrage in explaining the evolution of mutual fund alphas.

We focus on corporate bond funds for several reasons. One is simply that there are many of them, about 363 on average at any given time during our sample, and that the asset class that they represent is extremely large, until recently on par with the size of the U.S. equity market. Second, corporate bonds are conceptually somewhat close to equities, in the sense that there is tremendous diversity in potential holdings and that the underlying assets on which the securities lay claim are domestic firms. It is therefore conceivable that the skills required to generate alpha in one market could be valuable in the other. They are nevertheless a distinct asset class, whose returns are poorly

[13] Exceptions are the broker-sold indicator of Bergstresser et al. (2009), the active peer benchmarks approach of Hunter et al. (2014), and the regression R-squared of Amihud and Goyenko (2013), all of which were proposed in studies that analyzed both equity and bond funds.

explained by any combination of returns on Treasury bonds and the stock market index.

We factor-adjust fund returns using a 4-factor model that slightly expands the standard one in the literature. All factors are excess returns on Barclays (formerly Lehman) bond indexes and include the indexes for U.S. Treasuries, investment-grade corporates, high-yield corporates, and mortgage-backed securities. This departs from the standard model, introduced by Blake, Elton, and Gruber (1993) and used more recently by Huij and Derwall (2008), in which Treasuries and investment-grade corporate bonds are aggregated into a single factor. While the performance of these two classes of bonds was very similar for most of our sample, they markedly diverged from one another in the recent Financial Crisis. During this period, the returns on investment-grade bonds were not spanned by Treasuries and high yield, leading to our decision to include all three as separate factors.[14] As with Blake, Elton, and Gruber (1993), the mortgage-backed securities index is used to capture the forms of optionality specific to that asset class.

Our corporate bond fund sample is also from the CRSP Mutual Fund database, where funds are identified as focusing on corporate bonds if the CRSP style code (crsp_obj_cd) begins with "IC." Similar to the equity fund sample, we subject the corporate bond fund sample to screens to mitigate omission bias and incubation and back-fill bias. We again combine multiple share classes for each fund, focusing on the TNA-weighted aggregate share class.

An impediment to a full corporate bond fund analysis is that many predictors examined above cannot be computed because of data limitations. The chief limitation is that holdings of corporate bond funds are only available starting in September of 2003, which would lead to a sample that is too short for our purposes.[15] Individual fixed income security prices are also difficult to reconstruct prior to 2005, when reporting to the TRACE system became mandatory for all fixed income trades. As a result, our bond fund analysis considers only 13 predictors, 6 of which are measures of past performance.

Table 8 takes a first look at the performance of these predictors. Realized alphas are computed using rolling 36-month betas from the 4-factor model described above. From these alphas, we compute univariate quintile sorts for each of the 13 predictors along with average cross-sectional regression slopes. Our corporate bond fund data begin in January 1991, when monthly TNA data become available, so our first realized alphas are observed in January 1994. The sample ends in January 2015. The results tabulated use this entire 21-year period, without regard to the original sample period used in the study

[14] If we regress returns on the investment-grade corporate index on Treasuries and high- yield indexes, we obtain an R-squared of around 0.92 in the period from January 1990 to December 2006. In the period from January 2007 to December 2014, that R-squared drops to 0.76.

[15] Corporate bond holdings data are available from CRSP, but not the more standard Thompson-Reuters data set. Neither the CRSP MF data manual nor any study we are aware of provides sufficient information for us to judge the coverage and quality of the CRSP corporate bond mutual fund holdings data.

Table 8
Corporate bond fund alpha spreads

	Performance-based predictors												
	5-year alpha	1-year alpha	1-year return	Backtested alpha	P-S alpha	Active peer	R-squared	Expense ratio	Turnover	Fund size	Past flows	Direct sold	Abnormal cash
A. Average realized alphas from quintile sorts													
Low	−12.07	−14.94	−12.61	−16.11	−12.10	−10.32	−2.38	−3.39	−3.02	−4.76	−6.83	−6.81	−3.74
	(−4.41)	(−4.10)	(−3.88)	(−4.23)	(−4.11)	(−4.65)	(−0.83)	(−2.24)	(−1.91)	(−3.48)	(−3.25)	(−3.12)	(−1.44)
2	−6.33	−7.84	−6.46	−5.81	−5.71	−5.22	−5.61	−2.32	−5.26	−5.05	−5.22		−5.48
	(−3.86)	(−4.34)	(−3.44)	(−2.99)	(−3.29)	(−2.98)	(−2.62)	(−1.44)	(−2.75)	(−3.88)	(−3.08)		(−2.75)
3	−4.92	−4.20	−4.33	−3.55	−4.80	−3.01	−5.65	−4.08	−6.19	−5.49	−4.67		−5.41
	(−3.63)	(−3.53)	(−2.24)	(−1.81)	(−3.47)	(−1.65)	(−3.20)	(−2.49)	(−3.23)	(−3.65)	(−2.67)		(−2.17)
4	−2.23	−1.84	−2.22	−1.45	−2.15	−1.83	−6.16	−5.06	−4.66	−5.85	−3.85		−5.53
	(−1.81)	(−1.56)	(−1.45)	(−0.74)	(−1.88)	(−1.17)	(−3.70)	(−2.93)	(−2.66)	(−2.46)	(−2.39)		(−2.55)
High	−0.17	3.62	0.42	1.05	2.40	0.19	−5.32	−9.22	−4.89	−4.07	−3.62	−3.40	−4.74
	(−0.07)	(1.77)	(0.19)	(0.32)	(1.03)	(0.15)	(−4.94)	(−3.45)	(−2.54)	(−1.88)	(−2.24)	(−2.32)	(−2.51)
High–Low	11.90	18.56	13.03	17.17	14.50	10.51	−2.95	−5.83	−1.88	0.69	3.21	3.41	−1.00
	(6.83)	(5.34)	(3.42)	(5.31)	(6.32)	(5.30)	(−1.03)	(−2.59)	(−1.14)	(0.56)	(2.21)	(2.19)	(−0.53)
B. Average cross-sectional regression slopes													
	6.57	8.05	6.62	8.26	6.58	3.81	−1.48	−2.36	−0.28	0.20	2.12	1.65	−0.30
	(5.94)	(5.78)	(4.25)	(5.17)	(5.49)	(6.12)	(−1.90)	(−2.92)	(−0.54)	(0.45)	(2.13)	(2.19)	(−0.32)

This table examines the ability of different variables to predict returns on corporate bond funds. We consider all fund predictors that can be computed given available bond fund data, which excludes predictors based on portfolio holdings. In panel A we report average realized alphas (in basis points per month) for quintile portfolios. In panel B, we report average cross-sectional regression slope coefficients, in which realized alphas are regressed on a single predictor. t-statistics, in parentheses, are computed using Newey-West standard errors. The sample period is from 1994:01 to 2015:01.

that originally proposed each predictor, given that we view the entire history of corporate bond funds as being outside the original samples of the various studies.

Most of the 13 predictors already have, since their original studies, been applied to the corporate bond fund setting in other papers. Thus, our main contribution in this table is to reassess a variety of different predictors using standardized empirical methods.

In short, the table documents that all six performance persistence measures are highly significant in the corporate bond fund sample.[16] This finding is in line with Huij and Derwall (2008), who report similar results for some of the more straightforward performance measures.

We confirm the result of Blake, Elton, and Gruber (1993) that higher expense ratios are associated with lower alphas in corporate bond funds, as they are in equity funds. We also corroborate Chen and Qin (2016), who find evidence of the "smart money" effect in corporate bond funds. Finally, we find some evidence that higher *R*-squareds, computed based on our four bond market factors, are negatively associated with bond fund alphas. This effect is insignificant in quintile sorts and only marginally significant when implemented with cross-sectional regression. Our results thus only weakly confirm that the finding of Amihud and Goyenko (2013) holds for corporate bond funds as well.[17]

In contrast, we do not find any significant evidence of a turnover effect (Elton et al. 1993), a fund size effect (Chen et al. 2004), or a tendency of funds with high cash holdings to perform well (Simutin 2014) in bond funds, at least over the full sample period.

We do not believe that the lack of evidence for these three predictors indicates data snooping, as it is easy to see why scale and liquidity effects would be different for stock and bond funds. Rather, our view is that the significance of 9 or 10 out of 13 predictors is strong evidence that data snooping may not be as pervasive as has been claimed. At least in the case of mutual fund predictors, the data seem to be in conflict with the statement of Harvey, Liu, and Zhu (2016) that "most claimed research findings in financial economics are likely false."

Given our earlier results on the importance of arbitrage activity, a primary interest is in whether higher levels of arbitrage activity are associated with lower alpha spreads. An issue is that the arbitrage activity proxies considered previously, particularly short interest and turnover in the equity market, are

[16] When sorting on the basis of past alphas, we compute those alphas using the 4-factor model for bonds rather than the Fama-French-Carhart factors used in the equity fund analysis.

[17] Amihud and Goyenko (2013) themselves confirm their finding for corporate bond funds, but they employ a shorter sample period than ours and use different empirical specifications, whereas we apply common empirical methods for all predictors.

unlikely to be of much relevance for corporate bonds. We therefore consider a different set of arbitrage proxies that are more germane to this analysis.

The first measure we consider is the noise variable of Hu, Pan, and Wang (2013). This is a measure of the degree of nonsmoothness in the U.S. Treasury yield curve. The idea is that if capital is plentiful, then arbitrageurs will exploit small differences in the yields on bonds with slightly different maturities. As arbitrage capital is reduced, or as arbitrageurs become less willing to take risk, trading on the same differences becomes unattractive, with the result being that that differentials in yields of similar bonds persist.

The second measure we examine is the turnover in corporate bonds from SIFMA, which is computed similarly to the equity turnover measure used in earlier results on equity funds.

Our final measure is the average bid-ask spread for corporate bonds, computed by Adrian et al. (2015). Because the latter two series are based on TRACE data, they are only available starting in 2005.[18] As before, we will examine these arbitrage proxies in raw and detrended forms. We will also look at them individually and, after normalizing, as averages.[19] In this case, though, the shorter samples for corporate bond turnover and bid-ask spreads means that the composite arbitrage variable is the same as the Hu, Pan, and Wang noise measure through the end of 2004 and therefore may not be much more informative than the noise measure by itself.

We use these arbitrage activity proxies in fund-level panel regressions to determine whether a change in arbitrage activity is associated with any change in the degree of alpha generation by corporate bond funds. So that higher values of all arbitrage measures imply greater arbitrage activity, we multiply the Hu, Pan, and Wang (2013) measure and the average bid-ask spread by -1.

Table 9 presents the results, where regressions either include time and fund fixed effects or only time fixed effects, and where average scores are based on either percentiles or extreme quintiles.[20] Similar to the criteria applied to equity funds, we include a predictor in our analysis if its average full-sample alpha spread has a t-statistic above 1.5. We report regressions both with raw and detrended arbitrage proxies.

The regressions do not include out-of-sample effects for two reasons. First, it is not clear why there should be any out-of-sample effect in corporate bond funds given that most predictors were proposed based on their ability to forecast equity fund performance. Second, the corporate bond sample starts in 1991,

[18] We have also examined the Amihud (2002) ratio, which was also computed by Adrian et al. (2015). The results we obtain using this measure are consistent, both in magnitude and significance, with those reported in the tables.

[19] When we compute the composite arbitrage activity measure, the Hu, Pan, and Wang noise measure and the average bid-ask spread enter with a negative sign, because both are thought to be inversely related to arbitrage activity and/or liquidity. Corporate bond turnover enters with a positive sign.

[20] Because of data limitations, we do not examine the case with style-by-time fixed effects as in the equity funds setting. As described earlier, we lack data on bond prices and characteristics as well as on the holdings of bond funds, which would all be required to classify bond funds into styles.

Table 9
Arbitrage activity and corporate bond funds

	Fund & time fixed effects			Time fixed effects		
Arbitrage activity proxy (Z)	Avg. score (Avg$S$)	$Z\times$Avg$S$	Adj. R^2 (%) & # of obs.	Avg. score (Avg$S$)	$Z\times$Avg$S$	Adj. R^2 (%) & # of obs.
A. percentile scores						
Hu, Pan, & Wang noise	23.68	−14.23	8.21	27.19	−12.46	7.13
	(5.51)	(−4.14)	99,275	(6.47)	(−3.50)	96,863
Hu, Pan, & Wang noise	16.92	−13.48	8.20	21.42	−11.80	7.11
- detrended	(4.96)	(−4.06)	99,275	(7.25)	(−3.39)	96,863
Corp. bond turnover	−4.03	−10.66	8.56	4.32	−10.72	8.15
	(−0.76)	(−2.40)	43,548	(1.06)	(−2.22)	41,617
Corp. bond turnover - detrended	−9.60	−13.66	8.66	−0.40	−13.25	8.24
	(−1.42)	(−2.46)	43,548	(−0.07)	(−2.20)	41,617
Corp. bond bid-ask spread	30.57	−18.44	8.91	37.04	−16.35	8.44
	(3.19)	(−3.48)	43,548	(3.61)	(−2.71)	41,617
Corp. bond bid-ask spread	11.41	−17.54	8.88	20.29	−16.01	8.45
- detrended	(1.95)	(−3.65)	43,548	(4.32)	(−2.79)	41,617
Avg. of nondetrended arb. proxies	21.97	−16.69	8.08	25.81	−14.45	7.01
	(4.88)	(−3.59)	99,275	(5.81)	(−2.82)	96,863
Avg. of detrended arb. proxies	14.43	−14.48	8.05	19.33	−12.41	6.97
	(4.29)	(−3.37)	99,275	(7.08)	(−2.64)	96,863
B. Extreme quintile scores						
Hu, Pan, & Wang noise	21.37	−12.97	8.17	24.27	−11.37	7.04
	(5.46)	(−4.02)	99,275	(6.25)	(−3.40)	96,863
Hu, Pan, & Wang noise	15.26	−12.32	8.16	19.02	−10.79	7.03
- detrended	(4.97)	(−3.96)	99,275	(7.03)	(−3.30)	96,863
Corp. bond turnover	−4.17	−9.85	8.55	3.80	−9.76	8.11
	(−0.87)	(−2.45)	43,548	(1.00)	(−2.20)	41,617
Corp. bond turnover - detrended	−9.04	−12.54	8.64	−0.44	−12.06	8.19
	(−1.47)	(−2.47)	43,548	(−0.08)	(−2.18)	41,617
Corp. bond bid-ask spread	27.68	−16.86	8.88	33.92	−15.20	8.40
	(3.23)	(−3.44)	43,548	(3.62)	(−2.72)	41,617
Corp. bond bid-ask spread	10.11	−16.21	8.87	18.30	−14.91	8.41
- detrended	(1.97)	(−3.64)	43,548	(4.33)	(−2.81)	41,617
Avg. of nondetrended arb. proxies	19.92	−15.36	8.05	23.06	−13.25	6.93
	(4.89)	(−3.57)	99,275	(5.65)	(−2.79)	96,863
Avg. of detrended arb. proxies	12.99	−13.33	8.02	17.12	−11.38	6.90
	(4.31)	(−3.37)	99,275	(6.90)	(−2.62)	96,863

This table reports the results of fund-level panel regressions that analyze the relation between realized corporate bond fund alphas and various proxies for the level of bond market arbitrage activity. The regression equation is

$$\alpha_{it}=a+b\,\text{Avg}S_{it}+c\,Z_t\,\text{Avg}S_{it}+\epsilon_{it},$$

where α_{it} is the realized alpha in basis points for fund i in month t; Avg S_{it} is fund i's average percentile-based or extreme quintile-based score across all predictors in month t; and Z_t is an arbitrage activity proxy observable at the start of month t. Arbitrage proxies include the noise measure of Hu, Pan, and Wang (2013) multiplied by −1, the level of turnover in corporate bonds, and the average bid-ask spread in corporate bonds multiplied by −1. The sample is from 1994:01 to 2015:01 for the regressions with the noise measure and the average proxies. The sample starts in 2005:01 for the turnover and bid-ask spread regressions. For regressions with fund and time fixed effects, we cluster by date when computing standard errors. For regressions with only time fixed effects, we cluster by fund and date, include controls for expense ratio and log TNA, and exclude those two variables from the calculation of fund scores. t-statistics are in parentheses.

which is significantly later than the equity fund sample. This means that the in-sample period is very short or even nonexistent for many of our predictors.

The results in Table 9 are consistent in showing that proxies for greater arbitrage activity (less noise, greater turnover, and smaller spreads) are strongly associated with less performance predictability. The strongest results are for the

Hu, Pan, and Wang noise measure, though some of the significance of these results is likely due to the longer sample available for that measure. Regressions based on percentiles and extreme quintiles produce similar results.

In sum, the unconditional strength of most alpha predictors in the corporate bond fund sample shows that data snooping is unlikely to be the primary explanation for why those predictors were successful for equity funds in their original sample periods. The importance of arbitrage activity in shrinking the cross-section of fund alphas provides further evidence that time varying market efficiency is an important factor behind changes in fund performance.

5. Summary and Conclusions

Whether academic finance research is useful or not in practice largely rests on whether or not its findings continue to hold out of sample. In the case of variables shown in the finance literature to forecast future mutual fund alphas, we find that the out-of-sample performance is disappointing. Using several different econometric approaches, we show that much of the predictability in fund alphas disappears following the end of the sample.

We find that this decline is most likely the result of an increase in arbitrage activity over our sample period, with changing mutual fund competition perhaps playing a secondary role. The arbitrage activity variables we consider – aggregate measures of short interest, share turnover, and hedge fund AUM – have been shown by Chordia, Subrahmanyam, and Tong (2014) to be associated with shrinking alphas on asset pricing anomalies like size and momentum and may be proxies for the degree of market efficiency more generally. We find that these variables are important for mutual fund alphas as well and largely explain the decline in mutual fund alpha generation ability. After controlling for arbitrage, we find no evidence supporting data snooping or learning effects.

We also examine the performance of alpha predictors in a sample of corporate bond mutual funds. We find that most predictors, almost all of which were originally used to forecast equity fund performance, are also useful in forecasting the alphas of corporate bond funds. This performance also seems to change over time, however, with alpha spreads larger in times when arbitrage activity is lowest. Therefore, these results support the main findings of the paper, namely, that greater arbitrage activity appears to curtail the ability of fund managers to generate alpha.

The clear implication of our findings is that investment practitioners, who are known to use at least some of these measures to guide portfolio selection, may be engaging in an exercise that is of dubious relevance. More fundamentally, the dampening effect of greater arbitrage activity on mutual fund alpha could explain a number of empirical regularities, including the recent decline in mutual fund alphas (Barras, Scaillet, and Wermers 2010), the rise of hedge funds, and the trend toward indexation within the mutual fund industry. Finally, our results highlight the importance of allowing for time variation in return predictability. We believe that these issues merit greater attention than they have received thus far in the literature.

References

Adrian, T., M. Fleming, O. Shachar, and E. Vogt. 2015. Has U.S. corporate bond market liquidity deteriorated? *Liberty Street Economics*, October 5. libertystreeteconomics.newyorkfed.org/ 2015/10/has-us-corporate-bond-market-liquidity-deteriorated.html.

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

Amihud, Y., and R. Goyenko. 2013. Mutual fund's R^2 as predictor of performance. *Review of Financial Studies* 26:667–94.

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance* 65:179–216.

Bergstresser, D., J. Chalmers, and P. Tufano. 2009. Assessing the costs and benefits of brokers in the mutual fund industry. *Review of Financial Studies* 22:4129–56.

Berk, J. B., and R. Green. 2004. Mutual fund flows and performance in rational markets. *Journal of Political Economy* 112:1269–95.

Blake, C. R., E. J. Elton, and M. J. Gruber. 1993. The performance of bond mutual funds. *Journal of Business* 66:371–403.

Busse, J. A., and P. J. Irvine. 2006. Bayesian alphas and mutual fund persistence. *Journal of Finance* 61:2251–88.

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Chan, L. K. C., H. Chen, and J. Lakonishok. 2002. On mutual fund investment styles. *Review of Financial Studies* 15:1407–37.

Chen, J., H. Hong, M. Huang, and J. D. Kubik. 2004. Does fund size erode mutual fund performance? The role of liquidity and organization. *American Economic Review* 94:1276–302.

Chen, J., H. Hong, W. Jiang, and J. D. Kubik. 2013. Outsourcing mutual fund management: Firm boundaries, incentives, and performance. *Journal of Finance* 68:523–58.

Chen, Y., and W. Ferson. 2015. How many good and bad funds are there, really? Working Paper.

Chen, Y., and N. Qin. 2016. The behavior of investor flows in corporate bond mutual funds. *Management Science* 63:1365–81.

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

Christoffersen, S., and S. Sarkissian. 2009. City size and fund performance. *Journal of Financial Economics* 92:252–75.

Cohen, R. B., J. D. Coval, and L. Pástor. 2005. Judging fund managers by the company they keep. *Journal of Finance* 60:1057–96.

Cremers, M., and A. Petajisto. 2009. How active is your fund manager? A new measure that predicts performance. *Review of Financial Studies* 22:3329–65.

Doshi, H., R. Elkamhi, and M. Simutin. 2015. Managerial activeness and mutual fund performance. *Review of Asset Pricing Studies* 5:156–84.

Edelen, R. M., R. B. Evans, and G. B. Kadlec. 2007. Scale effects in mutual fund performance: The role of trading costs. Working Paper.

Elton, E. J., M. J. Gruber, and C. R. 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* 56:2415–30.

———. 2011. Holdings data, security returns, and the selection of superior mutual funds. *Journal of Financial and Quantitative Analysis* 46:341–67.

Elton, E. J., M. J. Gruber, S. Das, and M. Hlavka. 1993. Efficiency with costly information: A reinterpretation of evidence from managed portfolios. *Review of Financial Studies* 6:1–22.

Evans, R. B. 2010. Mutual fund incubation. *Journal of Finance* 65:1581–611.

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

———. 2010. Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance* 65:1915–47.

Gaspar, J. M., M. Massa, and P. Matos. 2006. Favoritism in mutual fund families? Evidence on strategic cross-fund subsidization. *Journal of Finance* 61:73–104.

Grinblatt, M., and S. Titman. 1994. A study of monthly mutual fund returns and performance evaluation techniques. *Journal of Financial and Quantitative Analysis* 29:419–44.

Grinblatt, M., S. Titman, and R. Wermers. 1995. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *American Economic Review* 85:1088–105.

Gruber, M. J. 1996. Another puzzle: The growth in actively managed mutual funds. *Journal of Finance* 51:783–810.

Gupta-Mukherjee, S. 2014. Investing in the "new economy": Mutual fund performance and the nature of the firm. *Journal of Financial and Quantitative Analysis* 49:165–91.

Harvey, C. R., and Y. Liu. 2018. Detecting repeatable performance. *Review of Financial Studies* 31:2499–552.

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

Hendricks, D., J. Patel, and R. Zeckhauser. 1993. Hot hands in mutual funds: Short-run persistence of relative performance, 1974-1988. *Journal of Finance* 48:93–130.

Hoberg, G., N. Kumar, and N. Prabhala. 2018. Mutual fund competition, managerial skill, and alpha persistence. *Review of Financial Studies* 31:1896–929.

Hu, G. X., J. Pan, and J. Wang. 2013. Noise as information for illiquidity. *Journal of Finance* 68:2341–82.

Huang, J., C. Sialm, and H. Zhang. 2011. Risk shifting and mutual fund performance. *Review of Financial Studies* 24:2575–616.

Huij, J., and J. Derwall. 2008. 'Hot hands' in bond funds. *Journal of Banking & Finance* 32:559–72.

Hunter, D., E. Kandel, S. Kandel, and R. Wermers. 2014. Mutual fund performance evaluation with active peer benchmarks. *Journal of Financial Economics* 112:1–29.

Ippolito, R. A. 1989. Efficiency with costly information: A study of mutual fund performance, 1965–1984. *Quarterly Journal of Economics* 104:1–23.

Jones, C. S., and L. Pomorski. 2016. Investing in disappearing anomalies. *Review of Finance* 21:237–67.

Jones, C. S., and J. Shanken. 2005. Mutual fund performance with learning across funds. *Journal of Financial Economics* 78:507–52.

Jylhä, P., and M. Suominen. 2011. Speculative capital and currency carry trades. *Journal of Financial Economics* 99:60–75.

Kacperczyk, M., and A. Seru. 2007. Fund manager use of public information: New evidence on managerial skills. *Journal of Finance* 62:485–528.

Kacperczyk, M., C. Sialm, and L. Zheng. 2005. On the industry concentration of actively managed equity mutual funds. *Journal of Finance* 60:1983–2011.

———. 2008. Unobserved actions of mutual funds. *Review of Financial Studies* 21:2379–416.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp. 2014. Time-varying fund manager skill. *Journal of Finance* 69:1455–84.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund ‘stars’ really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance* 61:2551–95.

Leamer, E. E. 1978. *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley & Sons.

Linnainmaa, J. T., and M. R. Roberts. 2018. The history of the cross-section of stock returns. *Review of Financial Studies* 31:2606–49.

Lo, A. W., and A. C. MacKinlay. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1:41–66.

———. 1990. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3:431–67.

Lou, D. 2012. A flow-based explanation for return predictability. *Review of Financial Studies* 25:3457–89.

Mamaysky, H., M. Spiegel, and H. Zhang. 2007. Improved forecasting of mutual fund alphas and betas. *Review of Finance* 11:359–400.

Massa, M. 2003. How do family strategies affect fund performance? When performance-maximization is not the only game in town. *Journal of Financial Economics* 67:249–304.

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

Nanda, V., Z. J. Wang, and L. Zheng. 2004. Family values and the star phenomenon: Strategies of mutual fund families. *Review of Financial Studies* 17:667–98.

Pástor, L., and R. F. Stambaugh. 2002. Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics* 63:315–49.

———. 2012. On the size of the active management industry. *Journal of Political Economy* 120:740–81.

Pástor, L., R. F. Stambaugh, and L. A. Taylor. 2015. Scale and skill in active management. *Journal of Financial Economics* 116:23–45.

Pollet, J. M., and M. Wilson. 2008. How does size affect mutual fund behavior? *Journal of Finance* 63:2941–69.

Schwert, G. W. 2003. Anomalies and market efficiency. In *Handbook of the economics of finance*, chap. 15, 939–74, M. Harris, G. Constantinides, and R. M. Stulz, eds. Amsterdam, the Netherlands: North-Holland.

Simutin, M. 2014. Cash holdings and mutual fund performance. *Review of Finance* 18:1425–64.

Wahal, S., and A. Y. Wang. 2011. Competition among mutual funds. *Journal of Financial Economics* 99:40–59.

Wermers, R. 2000. Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance* 55:1655–95.

White, H. 2000. A reality check for data snooping. *Econometrica* 68:1097–126.

Zheng, L. 1999. Is money smart? A study of mutual fund investors’ fund selection ability. *Journal of Finance* 54:901–33.