Skill and Efficiency in the U.S. Mutual Fund Industry

Dong Hwan Oh* and Andrew J. Patton†

September 3, 2025

Preliminary draft, please do not circulate.

Abstract

We propose a new measure of mutual fund manager ability: "efficiency" is the ability to accrue the risk premium associated with a risk factor. The familiar abnormal return, or alpha, is the sum of two distinct measures of ability: "efficiency" which is the weighted sum of the fund's (in)efficiencies across risk factors, and "skill," the component that is unrelated to factor exposures. Using a panel of U.S. equity mutual fund returns from 1999-2023, we document significant heterogeneity in mutual fund manager skill and efficiency. We employ regression trees and their extensions to capture this heterogeneity. We find that efficiency is more persistent than skill, and we show that future abnormal returns can be better predicted by decomposing lagged abnormal returns into skill and efficiency.

Keywords: mutual fund performance, trading costs, machine learning, heterogeneity.
J.E.L. Classification: G11, G12, G23, C58

*Federal Reserve Board. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors.

†Department of Economics, Duke University and School of Economics, Singapore Management University.

1 Introduction

U.S. mutual funds collectively manage more than $21 trillion in assets[1] and play an essential role in channeling household savings into equities, fixed income securities, and other asset classes. Given this scale, understanding whether and how mutual funds add value is a long-standing question in financial economics. Since Jensen (1968), academics have debated whether mutual fund managers possess genuine skill, or whether observed fund performances reflect exposure to systematic risk factors (Fama and French, 1993; Carhart, 1997) or luck (Barras *et al.*, 2010; Fama and French, 2010; Harvey and Liu, 2022).

Parallel to studies of manager skill, a growing literature (e.g., Keim and Madhavan, 1997; Wermers, 2000; Novy-Marx and Velikov, 2016) studies mutual funds' implementation costs, which detract from their abnormal returns. For example, Patton and Weller (2020) use Fama-MacBeth (1973) regressions to obtain an estimate of the "all-in" costs of implementing factor strategies, and show that, on average, mutual funds fail to fully capture the risk premia implied by factor models.

The revolution in how fund managers are evaluated, from raw returns to Sharpe (1964) or Treynor (1966) ratios, to the now-ubiquitous "risk-adjusted returns" (or alpha) of Jensen (1968) has immensely benefited investors, helping them to distinguish between risk premia and skill. But, given implementation costs, capacity constraints, and similar impediments faced by fund managers, what if accruing the full market risk premium on a portfolio is a non-trivial task? In such a world, one component of skill would be adding value beyond risk exposures; the traditional alpha. The other component would be accumulating the risk premium associated with those risk exposures; the component that enters the "risk-adjustment" term in alpha. This paper argues that both of these are a type of ability that a fund manager may possess, and shows that these abilities have distinct properties in the cross section and time series of U.S. mutual fund returns.

[1]Source: Federal Reserve Bank of St. Louis, series BOGZ1LM654090000Q.

We propose a decomposition of the familiar abnormal return, or alpha, into two components: "efficiency," which measures how well the manager is able to accrue the risk premia associated with a given risk factor, and "skill," the part of the fund's abnormal return that is unrelated to its exposures.[2] In conventional performance evaluation, full efficiency is assumed and so alpha is attributed entirely to skill. However, in the presence of implementation costs, accruing risk premia is itself a form of ability, and by measuring it we obtain important new insights into manager heterogeneity and performance persistence.

Using data on all U.S. equity mutual funds from 1999 to 2023, we document four main results. First, we show that there is statistically significant heterogeneity in both skill and efficiency across funds. Larger funds, for example, tend to be less efficient in harvesting premia while funds that are well explained by conventional factors ("high R^2" funds) tend to be highly efficient. These differences can be large: e.g., funds in the top quintile of R^2 earn 3.5% more per year per unit exposure to the market factor than funds in the bottom quintile. Thus the heteroegeneity in alpha documented in past work (e.g., Fama and French, 2010; Banegas *et al.*, 2013; Koijen, 2014) is also present in each of the components of alpha, skill and efficiency.

Second, we show that skill and efficiency are negatively related in the cross-section of funds: higher skilled funds tend to be less efficient, and vice versa. For example, funds in the top quintile of skill are three times more likely to be in the bottom quintile of efficiency than in the top quintile of efficiency. This is consistent with these tasks each requiring manager attention or effort, and good performance in one task detracts from performance in the other, see e.g., Kacperczyk *et al.* (2016).

Third, we find evidence that efficiency is more persistent than skill, with an autoregressive coefficient of 0.16 compared with 0.04. (Consistent with intuition, the persistence of alpha is between each of its two components, at 0.11.) This is consistent with efficiency

[2]Clearly, we use "efficiency" to describe a property that is related to, but distinct from, the mean-variance efficiency property of Markowitz (1952).

being a more fundamental property of the manager (or manager's investment process, liquidity management, execution costs, etc.) and with skill being more attributable to luck.

Finally, we show that forecasts of future abnormal returns can be significantly improved by decomposing past abnormal returns into skill and efficiency. Connecting with evidence that mutual fund performance persistence is mostly persistence in porly-performing funds (Brown and Goetzmann, 1995; Carhart, 1997), we find further gains in forecasts of future abnormal returns by decomposing skill and efficiency into positive and negative components. We find that positive skill and negative efficiency are key for predicting future abnormal returns.

We capture the heterogeneity in mutual fund efficiency using a variety of methods. Firstly, we use ex ante sorts of funds by observed characteristics like total net assets or the time series R^2 from the factor model (measuring how closely the fund is tracking factors). We also sort using the characteristics of the stocks held by the mutual funds (market capitalization, book-to-market ratio, and momentum), following the work of Daniel and Titman (1997). Given the preponderance of possible sort variables, we also consider regression trees (Breiman *et al.*, 1984, 2017), random forests (Breiman, 2001) and boosted regression trees (Friedman, 2001), to capture unknown forms of heterogeneity.

Our use of regression trees and their extensions connects this paper to the literature on machine learning methods in finance. Early work in this emerging area focused on machine learning methods for prediction (e.g., see Freyberger *et al.*, 2020; Gu *et al.*, 2020; Bianchi *et al.*, 2021; Aleti *et al.*, 2025; Kaniel *et al.*, 2023), and see Kelly and Xiu (2023) for a survey. In contrast, our use of these methods is focused on capturing unobserved heterogeneity through the lens of an economic model, linking this paper to the work of Bonhomme and Manresa (2015), Fan *et al.* (2022), Patton and Weller (2022) and Farrell *et al.* (2025).

The remainder of the paper is organized as follows. Section 2 describes our data and presents evidence that mutual funds, on average, fail to accrue all of the "on-paper" risk premia for the factors to which they are exposed. Section 3 develops our decomposition of abnormal returns into skill and efficiency and presents results on the cross-sectional heterogneiety in skill and efficiency. Section 4 shows how skill and efficiency are related to each other in the cross-section and the time series, and considers the predictive power of each for future abnormal returns. Section 5 concludes. Additional details and analyses are presented in the appendix.

2 Skill, efficiency, and abnormal returns

In this section we introduce the sample of mutual funds used in our analysis, and then show that mutual funds do not fully capture the risk premia associated with standard risk factors, i.e., are "inefficient" at accuring risk premia. Finally, we show that in the presence of inefficiencies, the famililar abnormal return is comprised of two components.

2.1 Data description

We use data on 5,046 U.S. domestic equity mutual funds available in CRSP, over the period 1999–2023. We follow Berk and van Binsbergen (2015), Pastor *et al.* (2015) and Patton and Weller (2020) for the data cleaning and filtering steps, which are described in detail in Appendix A. Summary statistics on the characteristics of these funds and their monthly returns is presented in Table 1. We see that approximately half of our full sample of funds (2,491) is available in each month of the sample, and that the average age of the funds in our sample is 14.6 years. The average (inflation-adjusted) total net assets (TNA) of mutual funds in our sample grew from \$1.8B to \$3.2B from the start to the end of our sample. The average TNA masks the well-known right skewness of mutual fund TNA, as shown by

Table 1: Summary statistics.
This table presents summary statistics on our sample of 5,046 United States domestic equity mutual funds, over the period 1999-2023. Panel A provides information on the time series of the number of active funds for each date as well as cross-sectional information on fund lifetimes and total net assets (TNA) at sample start, middle, and end (all measured in December 2023 dollars). Panel B reports distributional information on fund excess returns, such as the mean return, return volatility and the Sharpe ratio. ρ_{MKT} is the correlation with the S&P 500 index return, and R^2_{FF4} is the coefficient of determination from the Fama-French (1992) model augmented with the momentum factor of Carhart (1997).

Panel A: Fund characteristics

	Funds (number)	**Age (years)**	**TNA 1/99 ($m)**	**TNA 12/10 ($m)**	**TNA 12/23 ($m)**
Mean	2491.1	14.6	1794.6	1861.5	3208.1
Std Dev	261.3	10.0	6737.9	7595.6	12922.7
Q25	2285.5	6.4	71.6	72.3	86.6
Q50	2501.5	11.8	247.3	289.0	397.0
Q75	2698.0	21.3	953.2	1135.0	1579.7

Panel B: Fund return characteristics

	Mean Ret (% pa)	**Ret Vol (% pa)**	**Sharpe (Annualized)**	ρ_{MKT}	R^2_{FF4}
Mean	4.81	15.54	0.37	0.79	0.78
Std Dev	7.34	7.71	0.41	0.27	0.25
Q25	2.16	10.81	0.21	0.78	0.74
Q50	5.87	15.6	0.45	0.89	0.89
Q75	8.89	18.97	0.58	0.95	0.94

the cross-sectional quantiles of TNA. For example, in December 2023 the mean TNA was above the 75^{th} percentile of the distribution.

In Panel B of Table 1 we see that the average fund yielded 4.8% annualized average net-of-fees return, with annualized standard deviation of 15.5% and a Sharpe ratio of 0.37. Being domestic equity mutual funds, it is no surprise that they are highly correlated with the market, average correlation of 0.79, and well explained by the Fama-French (1992) Carhart (1997) four-factor model, with an average R^2 of 0.78, and a median R^2 of 0.89.

Figure 1: Number and average size of mutual funds, 1999-2023.
This figure shows the average total net assets (TNA) of the 5046 funds in our sample, measured in millions of December 2023 dollars, displayed on the left y-axis. On the right y-axis we show the average number of funds in our sample for each month.

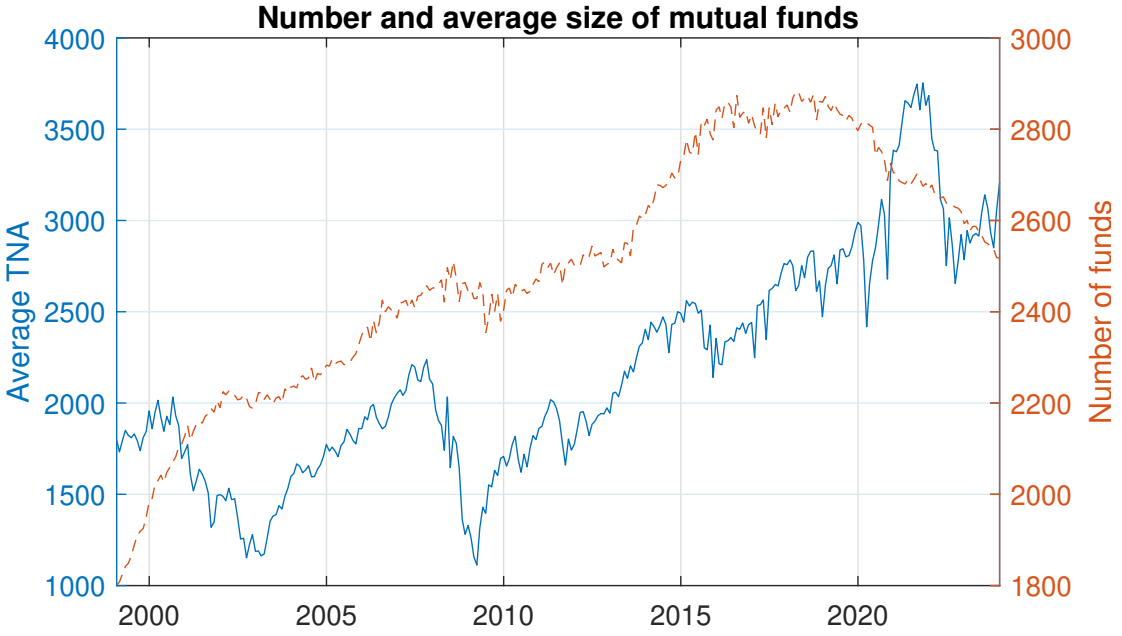


Figure 1 plots the average number of funds in each cross section in our sample period as well as the average TNA per fund. We observe that the number of funds has grown from just over 1,000 at the start of our sample, to around 2,900 in 2018, declining to about 2,500 at the end of our sample. The average size of mutual funds in our sample (in 2023 dollars) was around \$1.5B for the first decade of our sample, then rose strongly to over \$3.5B in 2021, declining to around \$2.6B in December 2023.

To construct the market equity (ME), book-to-market ratio (BM), and momentum (MO) variables at the mutual fund group level, we begin by obtaining quarterly mutual fund holdings data (including ticker, PERMNO, the number of shares held, and the market value of holdings) from the CRSP Mutual Funds Holdings database. For each stock identified by ticker and PERMNO, we collect monthly stock prices, returns, and shares outstanding from the CRSP monthly stock file. Additionally, we retrieve the quarterly accounting data

necessary for computing book value from the Compustat/CRSP merged database, using GVKEY and PERMNO as identifiers. Market equity (ME) is calculated as the product of a stock's price and the number of shares outstanding. Momentum (MO) is defined as the cumulative return over the past twelve months, excluding the most recent month (i.e., months $t-12$ to $t-2$). If the 12-month return is unavailable, we use the 6-month return as a proxy. The book-to-market ratio (BM) is computed monthly as book equity[3] divided by ME. After calculating monthly ME, BM, and MO for each stock, we aggregate these measures to the mutual fund level. Because CRSP holdings are reported quarterly, we assume that mutual funds maintain constant shareholdings between reporting dates. For each month, stock-level portfolio weights are calculated as the product of the number of shares held and the stock's month-end price, divided by the total value of all holdings in the mutual fund, i.e., the sum of these products across all stocks in the fund. This methodology follows established practices in Daniel *et al.* (1997), Ferson and Wang (2021), among others. Using these weights, mutual fund-level ME, BM, and MO are calculated as the monthly weighted averages of the corresponding stock-level characteristics. Finally, we aggregate to the mutual fund group level by computing TNA-weighted averages of each fund's characteristics, where TNA denotes the total net assets of the component mutual funds.

We adopt the popular four-factor model comprised of the market factor, a size factor ("SMB") and a value factor ("HML") following Fama and French (1992), augmented with the momentum factor of Carhart (1997). We obtain returns on these factors from Kenneth French's website.[4]

We use the procedure of Fama and MacBeth (1973) to estimate risk exposures and risk premia associated with the four factors. The first stage of this procedure is typically a time

[3]Book equity is assumed to remain constant within each quarter and is defined as the sum of Total Parent Stockholders' Equity (SEQQ) and Deferred Taxes and Investment Tax Credit (TXDITCQ), minus Preferred/Preference Stock (PSTKQ).

[4]https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

series regression using the last 36 or 60 months of returns, see Ferson (2019) and Kaniel *et al.* (2023) for example, which yields estimates of the risk exposures (betas) and abnormal returns (alphas) for each fund. However, since late 1998 daily returns on mutual funds have become available, facilitating the implementation of a superior method for measuring factor exposures: "realized betas" (Barndorff-Nielsen and Shephard, 2004). This method is used by Lewellen and Nagel (2006) for stock portfolios and by Akbas and Genc (2020) for mutual funds. We substitute the standard 36- or 60-month window of monthly returns with a three-month window (approximately 65 observations) of daily returns. This approach captures changes in funds' betas more quickly and facilitates more accurate estimates of risk premia in the second stage. If any of our mutual funds engage in factor timing (Bollen and Busse, 2001; Ammann *et al.*, 2020) our realized betas can capture this effect, in contrast with rolling-window estimates of beta using only monthly data. Details on how we compute realized betas for our sample of mutual funds are presented in Appendix B.

2.2 Do mutual funds acrrue the full risk premium?

We adopt the method described in Patton and Weller (2020) to answer the question posed in the section title: "on-paper" risk premia are estimated using a sample of 269 stock portfolios based on sorts on size, value, investment, etc.[5] We compare these with the risk premia estimates obtained estimating the same model on our sample of mutual funds. The difference between the "on-paper" risk premium and that achieved by mutual funds is the average (in)efficiency of mutual funds for that factor.

Table 2 reveals economically and statistically significant inefficiencies for the average mutual fund. The annualized on-paper market risk premium for our sample period is

[5]Specifically, we use portfolios based on one-way decile sorts on size, beta, book-to-market (BM), momentum, short-term reversal (STR), investment, and operating profitability (OP); two-way quintile sorts on size-BM, size-beta, size-momentum, size-STR, size-investment, and size-OP; and 49 industry portfolios formed based on SIC codes.

Table 2: Differences in earned risk premia.
This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B The first column uses 269 stock portfolios described in Section 2.1, and represents the "on-paper" risk premia for these factors. The second column presents estimates using our sample of 5,046 mutual funds, also described in Section 2.1. The third column presents the difference between the two estimates, and can be interpreted as the average "(in)efficiency" of mutual funds in our sample for a specific risk factor. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all four differences are zero. All t-statistics and p-values use Newey-West(1987) standard errors with five lags.

	Stocks	**Mutual funds**	**Difference**
MKT	8.970	5.526	-3.445
(t-stat)	(2.760)	(1.850)	(-4.623)
SMB	2.554	3.038	0.484
(t-stat)	(1.361)	(1.694)	(0.678)
HML	1.268	-0.909	-2.177
(t-stat)	(0.532)	(-0.343)	(-2.272)
MOM	1.693	0.348	-1.344
(t-stat)	(0.553)	(0.113)	(-1.215)
R^2	0.326	0.542	
$\bar{N}$	269	2,472	
Joint p-value			0.000

8.97%, while the premium earned by mutual funds for exposure to the market factor is only 5.53%, an efficiency loss of 3.4%, significant at the 0.01 level. The "value" factor (HML) risk premium is 1.27%, while mutual funds earn -0.91%, which is an effiency loss of -2.17%, significant at the 0.05 level. The other two factors (SMB and MOM) did not generate significant risk premia over our sample period, and mutual funds exhibit no inefficiency

in their exposures to those factors. A joint test that mutual funds are fully efficiency for all four factors rejects the null at the 0.01 level, indicating strong evidence that mutual funds do not fully capture the risk premia associated with the factors to which they are exposed.[6,7] This answers the question posed in the title of the section in the negative.

2.3 Decomposing the abnormal return

The canonical model for measuring the performance of mutual fund managers is:

$$\mathbb{E}[R_{i,t} - R_t^f] \quad = \quad \alpha_i + \boldsymbol{\beta_i}^\top \boldsymbol{\lambda} \tag{1}$$

where $\boldsymbol{\beta}_i$ measures the factor exposures of the fund, $\boldsymbol{\lambda}$ contains the risk premia for the factors, and α_i measures the abnormal return of the manager of fund i. However, the results in Table 2 reveal that mutual funds do not, in fact, earn the entire risk premium, $\boldsymbol{\lambda}$, for their exposures; instead, they typically earn less. If we denote the amount a given fund i actually earns as $\boldsymbol{\lambda} + \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ is typically negative, then this suggests an extension of the canonical model:

$$\mathbb{E}[R_{i,t} - R_t^f] \quad = \quad \psi_i + \boldsymbol{\beta_i}^\top (\boldsymbol{\lambda} + \boldsymbol{\eta_i}) \tag{2}$$

The intercept, ψ_i denotes the "skill" of the fund, the component of excess returns that is unrelated to its factor exposures. The (in)efficiency of the fund is captured by the vector $\boldsymbol{\eta}$, which measures how much, if any, of the risk premia associated with the factors,

[6] In Table A2 we re-do this analysis using monthly data to estimate betas and find qualitatively similar results: the significance of the inefficiency for the HML factor is lower, though it remains strong for the market factor, and the joint test of full efficiency for all four factors again rejects the null at the 0.01 level.

[7] The results in Table 2 use regressions that suppress the intercept in the stock portfolio regressions, thus imposing that these portfolios have zero alpha and overcoming collinearity with the vector of market betas. Mutual funds, on the other hand, may have alphas that differ from zero and so an intercept is included in those regressions. Table A3 in the appendix shows that when an intercept is also suppressed for mutual funds the inefficiencies are slightly smaller, but still strongly significant.

denoted by $\boldsymbol{\lambda}$, is lost. The aggregate efficiency of the fund is the weighted sum of the individual efficiencies, $\boldsymbol{\beta_i^\top \eta_i}$. Combining equations (1) and (2) reveals that, in the presence of inefficiency, the familiar abnormal return, α_i, is in fact comprised of *two* components: skill and the aggregate (in)efficiency:

$$\alpha_i \quad = \quad \psi_i + \Delta_i, \quad \text{where} \quad \Delta_i \equiv \boldsymbol{\beta_i^\top \eta_i} \tag{3}$$

If mutual funds are assumed to be fully efficient for all factor exposures, i.e. $\boldsymbol{\eta_i} = \mathbf{0} \; \forall \; i$, as is implicitly assumed in most of the literature to date, then equation (2) simplifies to equation (1) and we have $\psi_i = \alpha_i$. However, Table 2 presents strong evidence (significant at the 0.01 level) that $\boldsymbol{\eta_i} \neq \mathbf{0}$, and so both skill and efficiency each contribute to the abnormal return of a fund.

Equation (3) reveals another important feature for the analysis in this paper: skill and efficiency cannot be separately identified using only time series data on a single fund. The usual time series regression used to estimate factor exposures provides an estimate of abnormal returns, α_i, but without further information we cannot determine its constituent parts. In the next subsection we show how to use a panel of mutual funds to separately identify skill and efficiency, either via ex ante sorts of mutual funds or use machine learning methods.

3 Estimating mutual fund skill and efficiency

As discussed in the previous section, we cannot, without further assumptions, separately identify the skill and efficiency for a single mutual fund given only its returns. In this section we present a variety of methods for measuring the skill and efficiency of a panel of mutual funds.

3.1 Homogeneous factor efficiency

The simplest method is the one used to obtain Table 2, where we estimate the *average* efficiency for a cross-section of mutual funds. This method is optimal if all mutual funds are equally efficient for each factor. (We test, and reject, this restriction in the next section.)

Note that even though this approach assumes that all funds are equally efficient for each factor, the fact that average efficiency differs across factors (e.g., the average inefficiency is -3.4% for the market factor, -2.2% for the value factor, and not different from zero for the size and momentum factors) and factor exposures vary across mutual funds, the aggregate (in)efficiency will vary across funds:

$$\Delta_{i,t} = \boldsymbol{\beta}_{i,t-1}^{\top} \bar{\boldsymbol{\eta}}_t \tag{4}$$

Thus even this simple specification we can capture cross-sectional variation in efficiency arising from cross-sectional variations in factor loadings.

If accruing all of the risk premia associated with a set of risk factors is difficult, and Table 2 shows that the average mutual fund fails to do so, this raises the question of whether some mutual funds are able to do it better than others, i.e., are some mutual funds more efficient than others?

3.2 Are all mutual funds equally good at accruing risk premia?

We firstly allow for heterogeneity in efficiency by grouping funds by observable characteristics and examining the efficiency of the funds in each group separately. This imposes within-group homogeneity but allows for cross-group heterogeneity, thereby generalizing the approch presented in the previous section.

As discussed in Section 2.1, there are vast differences in the total net assets (TNA) across funds: for example, in the last month of our sample the TNA of a fund at the 75^{th}

Table 3: Heterogeneity in mutual fund efficiency: TNA sorts.
This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, as well as abnormal returns in the first row, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B. The first column uses all mutual funds, and corresponds to the second column in Table 2. The next five columns report parameter estimates separately for each quintile of mutual funds sorted by total net assets. The last two columns present the p-values (asymptotic and simulation-based) associated with a test that the parameter estimates in that row are equal across all quintiles. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The row labeled $\overline{TNA}$ reports the average total net assets (in millions of December 2023 dollars) of the funds in that column. The last row reports the p-value on the joint test that all quintiles have the same parameters. All t-statistics and asymptotic p-values use Newey-West(1987) standard errors with five lags.

		Total Net Asset Quntile					***p*-values**	
	All funds	**Small**	**2**	**3**	**4**	**Large**	**Asymp.**	**Sim.**
MKT	5.526	6.574	5.396	4.983	4.640	4.844	0.091	0.101
(t-stat)	(1.850)	(2.213)	(1.799)	(1.614)	(1.541)	(1.610)		
SMB	3.038	2.966	3.105	3.152	3.614	2.721	0.263	0.284
(t-stat)	(1.694)	(1.702)	(1.719)	(1.662)	(1.942)	(1.472)		
HML	-0.909	-1.887	-0.016	-0.285	-1.035	-0.437	0.035	0.039
(t-stat)	(-0.343)	(-0.722)	(-0.006)	(-0.102)	(-0.376)	(-0.156)		
MOM	0.348	0.102	-0.522	1.659	-0.429	0.643	0.055	0.082
(t-stat)	(0.113)	(0.034)	(-0.162)	(0.522)	(-0.129)	(0.195)		
Const	1.446	0.437	1.545	1.899	2.032	2.086	0.141	0.170
(t-stat)	(1.890)	(0.625)	(2.041)	(2.378)	(2.122)	(2.174)		
R^2	0.542	0.529	0.555	0.571	0.572	0.595		
$\bar{N}$	2,472	494	495	494	495	494		
$\overline{TNA}$	2,146	25	98	299	937	9,377		
Joint p-val							0.000	0.010

percentile is over 18 times as large as the fund at the 25^{th} percentile, and for the 90^{th} and 10^{th} percentiles the ratio is nearly 250. This motivates Table 3, which sorts mutual funds into quintiles based on their TNA at the end of the previous month. We present tests of equal efficiency across quintile groups using both the usual asymptotic χ^2 p-values and p-values constructed using a simulation-based approach described in detail in Appendix C.

Table 3 reveals significant heterogeneity in efficiency across TNA quintiles. The efficiency for the market factor, in the first row, is almost monotonically declining in TNA, going from 6.6% for the smallest funds to 4.8% for the largest. The p-value that these five estimates are equal is around 0.10 using either the asymptotic or the simulation-based approaches. For value (HML) we reject the null of equal efficiency at the 0.05 level, and for momentum we reject at the 0.10 level. The average skill for each quintile, captured via the intercept reported in the bottom row of the top panel, is monotonically increasing in TNA. Combining all coefficients and testing whether all quintiles share the same parameters we strongly rejects the null, with a p-value of 0.01. Not all funds are equally skilled or efficient.

Next we consider sorting funds by their time series R^2 from the Fama-French (1992)-Carhart (1997) model. Recall from Table 1 that the median fund has an R^2 of 0.89, revealing that these four factors explain most of the variation in fund returns, on average. However, there is significant cross-sectional variation in R^2, with the lowest quintile having an average R^2 of around one-third, while the top quintile has an average R^2 of over 97%. Clearly some funds hew very closely to these four factors, while others take on more idiosyncratic risks.

Table 4 provides further evidence of heterogeneity in skill and efficiency across mutual funds. High R^2 funds tend to be better at earning the market risk premium than low R^2 funds: 7.7% versus 4.2%, and the test that efficiency is equal across all quintiles rejects the null at the 0.10 level. High R^2 funds are also better at earning the size risk premium: 2.3% versus 1.6%, though the null of homogeneity is not rejected for this factor. Table 4 also reveals important heterogeneity in skill, the component of abnormal returns that is

Table 4: Heterogeneity in mutual fund efficiency: R^2 sorts.
This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, as well as abnormal returns in the first row, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B. The first column uses all mutual funds, and corresponds to the second column in Table 2. The next five columns report parameter estimates separately for each quintile of mutual funds sorted by their time series R^2 from the Fama-French (1992)-Carhart (1997) model. The last two columns present the p-values (asymptotic and simulation-based) associated with a test that the parameter estimates in that row are equal across all quintiles. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The row labeled $\overline{R^2}_{FF4}$ reports the average R^2 from the French (1992)-Carhart (1997) model for the funds in that column. The last row reports the p-value on the joint test that all quintiles have the same parameters. All t-statistics and asymptotic p-values use Newey-West(1987) standard errors with five lags.

		R^2_{FF4} **Quintile**					p**-values**	
	All funds	**Low**	**2**	**3**	**4**	**High**	**Asymp.**	**Sim.**
MKT	5.526	4.158	7	7.588	5.962	7.687	0.063	0.094
(t-stat)	(1.85)	(1.169)	(2.295)	(2.323)	(1.928)	(2.538)		
SMB	3.038	1.630	2.649	2.213	2.894	2.304	0.300	0.323
(t-stat)	(1.694)	(0.765)	(1.444)	(1.232)	(1.513)	(1.257)		
HML	-0.909	-0.617	-1.263	-0.264	0.476	0.550	0.367	0.428
(t-stat)	(-0.343)	(-0.217)	(-0.452)	(-0.096)	(0.169)	(0.199)		
MOM	0.348	1.964	0.476	-0.387	0.539	1.287	0.698	0.743
(t-stat)	(0.113)	(0.577)	(0.158)	(-0.118)	(0.159)	(0.366)		
Const	1.446	2.180	0.714	-0.253	0.720	-1.038	0.000	0.000
(t-stat)	(1.890)	(2.868)	(0.729)	(-0.224)	(0.842)	(-1.594)		
R^2	0.542	0.439	0.489	0.554	0.598	0.738		
$\bar{N}$	2,472	494	495	494	495	494		
$\bar{R}^2_{FF4}$	0.803	0.340	0.832	0.917	0.949	0.977		
Joint p-val							0.000	0.007

unrelated to factor exposures: low R^2 funds have much higher skill than high R^2 funds, 2.2% compared with -1.0%, and a test that skill is equal across all R^2 quintiles rejects the null at the 0.01 level. The results in Table 4 are consistent with economic intuition: funds that track the four factors closely should aim to capture all, or most, of the risk premia associated with those factors, while funds that pursue more idiosyncratic strategies pay less attention to the risk premia of these factors, and earn their abnormal returns via activities orthogonal to their risk exposures.[8]

We next present results when mutual funds are sorted based on the value-weighted characteristics of the stocks in their portfolios. Like Daniel *et al.* (1997) and Ferson and Wang (2021), we focus on three stock characteristics: market capitalization, book-to-market ratio, and momentum. Table 5 presents results using the market capitalization sort and reveals that funds that invest in larger-cap stocks tend to be better at accruing the MKT risk premium, with the estimates ranging from 3.9% for the smallest quintile to 6.2% for the largest. The strongest source of heterogeneity across quintiles is average skill, with an estimate of 3.6% for funds with the smallest market-cap stocks in their portfolios, and 0.7% for funds with the largest-cap stocks. A joint test of equal parameters across all quintiles rejects the null at the 0.01 level.

Table 6 presents results sorting funds by the value/growth characteristic of the stocks in their portfolios. We find that funds in the middle quintile, so neither "value" nor "growth" focused, are the most efficient for the MKT factor, earning 5.7%, compared with 4.8& for growth funds and 3.4% for value funds, and we reject the null of equal efficiency for MKT at the 0.10 level. Interestingly, the middle quintile is the least efficient at earning the HML premium, with an estimate of -2.7%, compared with -1.6% for growth funds and 0.9% for value funds. We reject the null of equal efficiency for the HML factor at the 0.05 level

[8] Amihud and Goyenko (2013) consider the time series R^2 as a fund characteristic and find funds with lower R^2 have higher abnormal returns than funds that more closely track systematic factors. The results in Table 4 can be interpreted as a decomposition of their result.

Table 5: Heterogeneity in mutual fund efficiency: Sorts by ME of holdings. This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, as well as abnormal returns in the first row, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B. The first column uses all mutual funds for which we have holdings data, and differs slightly from the second column in Table 2 due to missing data. The next five columns report parameter estimates separately for each quintile of mutual funds sorted by the value-weighted average of the market capitalization (ME) of their stock holdings. The last two columns present the p-values (asymptotic and simulation-based) associated with a test that the parameter estimates in that row are equal across all quintiles. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all quintiles have the same parameters. All t-statistics and asymptotic p-values use Newey-West(1987) standard errors with five lags.

		ME Quintile					p**-vals**	
	All funds	**Small**	**2**	**3**	**4**	**Large**	**Asymp.**	**Sim.**
MKT	5.526	3.938	5.720	5.634	6.177	6.233	0.104	0.135
(t-stat)	(1.850)	(1.368)	(1.965)	(1.774)	(2.078)	(2.020)		
SMB	3.038	2.814	1.765	2.835	0.673	3.305	0.185	0.213
(t-stat)	(1.694)	(1.516)	(0.959)	(1.832)	(0.353)	(1.465)		
HML	-0.909	0.676	-1.483	-1.733	-1.853	-0.879	0.264	0.301
(t-stat)	(-0.343)	(0.223)	(-0.533)	(-0.621)	(-0.718)	(-0.326)		
MOM	0.348	-0.220	0.645	0.744	1.134	0.413	0.896	0.908
(t-stat)	(0.113)	(-0.061)	(0.204)	(0.238)	(0.388)	(0.138)		
Const	1.446	3.582	1.832	1.084	0.800	0.669	0.002	0.002
(t-stat)	(1.890)	(3.158)	(1.874)	(1.612)	(0.896)	(0.958)		
R^2	0.542	0.501	0.478	0.632	0.498	0.553		
$\bar{N}$	2,472	494	495	494	495	494		
Joint p-val							0.000	0.009

Table 6: Heterogeneity in mutual fund efficiency: Sorts by BM of holdings. This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, as well as abnormal returns in the first row, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B. The first column uses all mutual funds for which we have holdings data, and differs slightly from the second column in Table 2 due to missing data. The next five columns report parameter estimates separately for each quintile of mutual funds sorted by the value-weighted average of the book-to-market ratio (BM) of their stock holdings. The last two columns present the p-values (asymptotic and simulation-based) associated with a test that the parameter estimates in that row are equal across all quintiles. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all quintiles have the same parameters. All t-statistics and asymptotic p-values use Newey-West(1987) standard errors with five lags.

		BM Quintile					p**-vals**	
	All funds	**Low**	**2**	**3**	**4**	**High**	**Asymp.**	**Sim.**
MKT	5.526	4.770	5.384	5.725	4.840	3.448	0.082	0.093
(t-stat)	(1.850)	(1.602)	(1.769)	(1.830)	(1.662)	(1.230)		
SMB	3.038	1.689	2.145	3.373	3.167	3.410	0.533	0.576
(t-stat)	(1.694)	(0.847)	(1.102)	(2.178)	(1.824)	(2.009)		
HML	-0.909	-1.643	-2.380	-2.745	-0.062	0.912	0.042	0.062
(t-stat)	(-0.343)	(-0.671)	(-0.995)	(-0.870)	(-0.025)	(0.324)		
MOM	0.348	-0.035	0.173	-0.157	0.275	0.245	0.999	0.998
(t-stat)	(0.113)	(-0.012)	(0.054)	(-0.048)	(0.094)	(0.073)		
Const	1.446	1.938	1.576	0.979	1.913	3.170	0.033	0.042
(t-stat)	(1.890)	(1.776)	(2.099)	(1.473)	(2.086)	(2.831)		
R^2	0.542	0.481	0.579	0.638	0.499	0.453		
$\bar{N}$	2,472	494	495	494	495	494		
Joint p-val							0.000	0.010

using the asymptotic test and the 0.10 level using the simulation-based test. We also see a U-shaped pattern in average skill, with the middle quintile having skill of 1.0% compared with 3.2% for the value quintile. The joint test rejects the null across all parameters at the 0.01 level.

We present the results based on sorts using the MOM characteristic of their holdings in Table A4 in the Appendix. Using the usual asymptotic p-value we reject the null of equal efficiency across quintile portfolios at the 0.05 level, however the simulation-based p-value is 0.17 and thus is far from rejecting the null of equal efficiency. Given the size distortion of the joint test reported in Appendix C, we focus on the simulation-based p-value rather than the asymptotic p-value, and conclude that we do not have evidence to reject the null when using momentum as a sorting variable.

The tables in this section answer the question in the posed in the title firmly in the negative: accruing the "on-paper" risk premium associated with risk factors is difficult (Table 2) and funds vary significantly in their ability to do so (Tables 3 to 6).

The abundance of variables available for ex ante sorting immediately raises the further question of whether one variable is more powerful than the others, and whether a combination of variables could perform better than any single variable. In the next section we adopt a popular method from machine learning to address these questions.

3.3 Regression trees for learning about skill and efficiency

The analysis in the previous section relies the selected sorting variable (e.g., TNA or R^2_{FF4}) and thresholds (quintiles, in the previous section) being a useful way of grouping mutual funds. An alternative to selecting the sorting variable and thresholds ex ante is to let the data determine the optimal grouping of funds. "Regression trees" (Breiman *et al.*, 1984, Breiman *et al.*, 2017) are perfectly suited to this task. Given a set of state variables, possibly a large number, a regression trees sequentially select the "splitting" variable and

the split threshold that maximizes the fit of the overall model. The procedure continues to split the data until the number of splits has reached a pre-determined maximum, or until another constraint (e.g, a constraint on the minimum number of observations in each split) is binding. Further details on the implementation of regression trees and their extensions are presented in Appendix D.

Regression trees naturally nest the ex ante splits considered in the previous section: the tree could select the same splitting variable (e.g., TNA) for all splits, and could select the quintiles of that variable as thresholds. However, with a variety of spitting variables to choose from, and flexibility on the threshold for each variable, such an outcome is unlikely.

We consider a total of 16 state variables across four categories: past performance (fund returns over the past 3 and 12 months, and cross-sectional ranks of fund returns over the past 3 and 12 months); flows (net flows into/out of the fund over the past 1, 3, and 12 months, and cross-sectional ranks of these flows); fund characteristics (cross-sectional ranks of TNA, R^2_{FF4}, and R^2_{CAPM}); fund holdings characteristics (cross-sectional ranks of the value-weighted average of the stocks in each fund's portfolio, according to size, value and momentum). We consider a grid of 19 thresholds for each of these variables, based on their 0.05, 0.1, ..., 0.95 percentiles.

It is possible to estimate a separate tree for each cross-section in our sample, however this can lead to over-fitting, as the tree structure changes more frequently than is economically plausible. To address this, we impose that the tree structure is the same for each calendar year.[9] This allows for low-frequency changes in the tree, and improves estimation accuracy by using 12 months of data to estimate the structure of the tree.

Regression trees, like other machine learning methods, over-fit in-sample and require regularization. We use five-fold cross validation to choose the number of splits in the tree as follows: we randomly assign funds into five equally-sized groups, and then use four of these

[9]Naturally we allow the regression parameters to vary each month, capturing variation in realized factor returns.

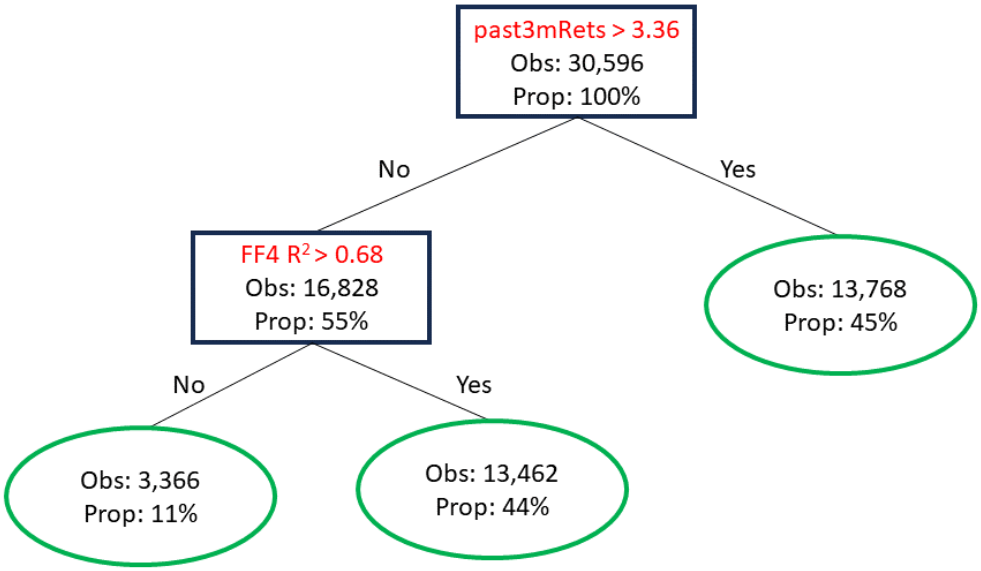
groups estimate the tree structure with number of leaves ranging from one (corresponding to no splits at all) up to ten, and finally we evaluate the fit on the fifth group. We repeat this so that each one of the groups is used as the out-of-sample group, and we sum the out-of-sample fit across all five "folds." We choose the number of leaves that maximizes the out-of-sample fit. Table A5 in Appendix E reports the number of leaves selected for each year. The optimal number of leaves ranges from one (where no tree structure is used, and the model reduces to simple OLS) to nine, and the average number of leaves is 4.4.

As an example, Figure 2 presents the optimal tree for 2023. The optimal number of leaves in that year is three. The first split separates funds approximately equally (55% and 45%) based on their returns over the past three months, and the second split separates the low-performing funds into low R^2_{FF4} funds and the rest (11% and 44% of the total number of funds).

We also consider random forests (Breiman, 2001), a popular extension of regression trees, to capture heterogeneity in mutual fund efficiency. Random forests have been found to produce forecasts with lower variance, and have greater robustness to noise. Using the optimal number of leaves from the cross-validation of the regression tree, we implement the random forest using 500 bootstrap samples of the original data, obtaining the regression tree forecast on each sample, and then averaging the resulting forecasts.

Finally, we consider boosted regression trees (BRTs) (Friedman, 2001) as an alternative means of capturing heterogeneity. BRTs improve on regression trees by sequentially modeling the residuals from the previous step, and can capture nonlinearities that might be missed in a regression tree. BRTs have a second hyperparameter, beyond the number of leaves in each tree, in the number of boosting iterations. We use the same five-fold cross-validation procedure described above to jointly tune the number of leaves and the number of boosting iterations. We allow up to ten leaves in the tree, and up to 25 boosting

Figure 2: Regression tree for 2023.
This figure describes the optimal tree for the last year of our sample period, 2023. The number of leaves is determined by five-fold cross-validation. The sample is split using the state variable and threshold given in the rectangles. The number of fund-months in each "leaf," as well as the proportion of the sample it represents, is reported.



iterations.[10]

Table 7 compares the cross-sectional R^2 from the eight methods for capturing variation in mutual fund efficiency. We show the average across all 300 months, as well as the 25^{th}, 50^{th}, and 75^{th} percentiles. Panel A shows that the average fit using standard OLS is 54%. This rises to 56% when splitting funds by TNA quintiles, and rises further to 59% when splitting funds by R^2_{FF4} quintiles. Sorting by fund holdings characteristics ME and BM also lead to R^2s of (just under) 59%. The regression tree and extensions have the highest average fits, ranging from 60% to 62%. Note that the tree achieves this a better fit with fewer parameters than the quintile sorts, as the latter estimate a total of 25 parameters (four risk premia and an intercept, for each of the five quintiles), while the tree model has

[10]We set the learning rate for the BRT at 0.05, and do not additionally tune this parameter.

Table 7: Cross-sectional fits.
This table presents summary statistics on the R^2 from Fama-MacBeth (1973) second-stage regressions for mutual fund returns. These regressions are run each month and this table presents the mean and three quartiles of these values across the 300 months in our sample. The first column presents results for the standard model, using OLS on the entire sample of mutual funds. The next four columns use OLS separately on quintiles of funds sorted by total net assets, their time series R^2 from the Fama-French (1991)-Carhart (1997) four-factor model, or the ME and BM characteristics of their stock holdings. The last three columns present results using the regression tree, random forest, and boosted regression tree (BRT) described in Section 3.3. Panel A presents results using realized betas, which are constructed using daily data in the preceding three months, and Panel B presents results using betas estimated using a 36-month window of returns.

		Ex-ante quintile sorts						
	OLS	**TNA**	R^2_{FF4}	**ME**	**BM**	**Tree**	**Forest**	**BRT**
Panel A: Realized betas								
Mean	0.541	0.559	0.594	0.589	0.585	0.604	0.612	0.621
Q25	0.395	0.417	0.463	0.453	0.457	0.482	0.503	0.503
Q50	0.569	0.579	0.618	0.609	0.609	0.637	0.645	0.649
Q75	0.707	0.716	0.746	0.742	0.743	0.754	0.764	0.766
Panel B: Monthly betas								
Mean	0.514	0.531	0.562	0.564	0.560	0.579	0.595	0.599
Q25	0.355	0.378	0.423	0.425	0.424	0.438	0.462	0.466
Q50	0.532	0.545	0.594	0.585	0.572	0.620	0.631	0.639
Q75	0.680	0.697	0.712	0.719	0.714	0.728	0.738	0.740

an average of 4.36 leaves, and so is based on an average of 21.8 parameters. It is noteworthy, though, that the machine learning-based methods, when regularized using cross validation, do not have dramatically higher fits than the best ex-ante sorts, indicating that these sorts capture much of the heteregeneity in the sample.

Panel B of Table 7 presents results when using monthly betas rather than realized betas. We see that the improvements in fit across the cross-sectional models for efficiency

are qualitatively the same for these betas. Comparing the average R^2s across the two panels, we observe that using realized betas yields R^2s that are around 2.5% higher than when using monthly betas in the same model, consistent with the former being more accurate measures of mutual fund factor exposures.

4 Decomposing mutual fund returns

In this section we decompose mutual fund returns into components related to risk premia, efficiency and skill. Recall the proposed model for mutual fund returns:

$$R_{i,t} = R_t^f + \psi_i + \boldsymbol{\beta}_{i,t-1}^\top(\boldsymbol{\lambda}_t + \boldsymbol{\eta}_{i,t}) + \varepsilon_{i,t} \tag{5}$$

We estimate "on-paper" risk premia using the Fama-MacBeth (1973) procedure on a panel of 269 stock portfolios, as displayed in Table 2, and denote these as $\boldsymbol{\lambda}_t^{Stocks}$. The abnormal return is estimated as the difference between the average return on the fund and the average "normal" return on the fund, equal to the risk free rate plus the risk premia associated with the fund's factor exposures:

$$\alpha_i = \mathbb{E}\left[R_{i,t}\right] - \mathbb{E}\left[R_t^f + \boldsymbol{\beta}_{i,t-1}^\top \boldsymbol{\lambda}_t^{Stock}\right] \tag{6}$$

We estimate $\boldsymbol{\lambda}_{i,t}$ using one of the four methods presented in Section 3, and from those we obtain the efficiency for each fund in our sample:

$$\boldsymbol{\eta}_{i,t} = \boldsymbol{\lambda}_{i,t} - \boldsymbol{\lambda}_t^{Stocks} \tag{7}$$

Finally, we estimate skill as

$$\psi_i = \alpha_i - \Delta_i \tag{8}$$

$$\text{where} \quad \Delta_i \equiv \mathbb{E}[\boldsymbol{\beta}_{i,t-1}^{\top} \boldsymbol{\eta}_{i,t}] \tag{9}$$

is the average aggregate efficiency of fund i.

We now investigate the properties of these components in the cross-section of mutual funds, and show that this finer decomposition of mutual fund returns improves our understanding of mutual fund performance.

4.1 Skill vs. Efficiency

A key motivation for this paper is that in the presence of inefficiencies in accruing risk premia, the familiar abnormal return is the sum of two terms: the average aggregate (in)efficiency, $\Delta_i \equiv \mathbb{E}[\boldsymbol{\beta}_{i,t-1}^{\top} \boldsymbol{\eta}_{i,t}]$, and a "skill" component unrelated to factor exposures, ψ_i. Figure 3 plots average aggregate efficiency against skill for all funds in our sample, using the tree model to capture heterogeneity. The red circle in Figure 3 denotes the average skill (0.76%) and efficiency (-2.34%) across all funds and all periods. Most interestingly, Figure 3 reveals that these measures of fund manager ability are negatively related in the cross section: the correlation is -0.21, and the OLS fitted line has slope -0.18 with a t-statistic of -14.3.[11] Thus some mutual funds achieve their abnormal returns by increased efficiency, while others do so via skill. Relatively few exhibit both positive efficiency and positive skill (and thus lying in the north-east quadrant of Figure 3).

Next we investigate the statistical significance of these measures of fund manager ability.

[11] If the estimates of individual mutual fund efficiency were pure noise, uncorrelated in the cross section, then the observed skill-efficiency relationship would also be negative, with correlation equal to $-\sqrt{V[\Delta_i]/V[\psi_i]}$, which equals -0.86 in our sample, compared with the actual correlation of -0.21. The difference arises because our estimates of efficiency are positively correlated with abnormal returns, with correlation of 0.55, despite being computed in a completely different step (cross-sectional rather than time series), consistent with the efficiency estimates not being pure noise.

Figure 3: Skill and efficiency of US mutual funds.
This figure plots skill (ψ_i) and aggregate efficiency ($\boldsymbol{\beta_i}^\top \boldsymbol{\eta_i}$), measured in annualized percent, for all mutual funds in our sample, averaged over all months for which they are in the sample. The thin dashed line denotes the "zero abnormal return" line, where skill and efficiency sum to zero. The thick dashed line is the OLS regression line. The red circle denotes the average skill and efficiency across all funds and all periods.

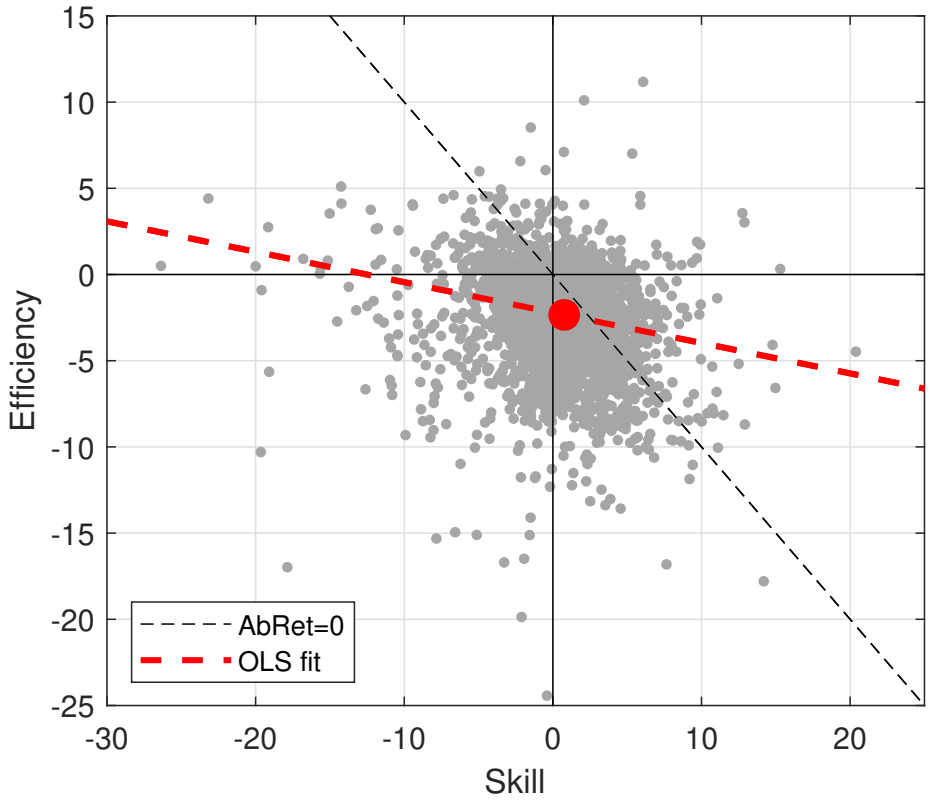


Table 8 presents a two-way contingency table, categorizing funds as having negative, zero, or positive skill, and negative, full, or positive efficiency, by testing these two measures are equal to zero at the 0.10 level. The right column of Table 8 presents the proportion of funds with positive, zero, and negative abnormal returns. We see that nearly half of funds (45.3%) are statistically fully efficient, a negligible fraction (2.4%, when the level of the test is 10%) are positively efficient, and slightly over half of funds are significantly inefficient. In contrast, around three-quarters of funds are zero skill, about a fifth exhibit positive skill, and a negligible proportion are negatively skilled. Just over half of funds live on the anti-diagonal of this contingency table, consistent with the negative correlation

Table 8: Skill, efficiency, and abnormal returns: Statistical significance.
This table presents results on the proportion of mutual funds that have skill, efficiency, and abnormal returns that are significantly different from zero at the 10% level. Estimates of skill and efficiency are based on the Tree model described in Section 3.3. "Full efficiency" corresponds to an estimate that is not different from zero. The left panel shows a two-way contingency table for skill and efficiency, and the right column shows the results for abnormal returns.

		Efficiency				***Abnormal Return***
		Negative	**Full**	**Positive**	**Total**	
Skill	**Negative**	0.006	0.024	0.002	0.033	0.202
	Zero	0.380	0.366	0.013	0.759	0.754
	Positive	0.137	0.062	0.009	0.209	0.045
	Total	0.523	0.453	0.024	1	1

between these measures of ability presented in Figure 3. The modal cell in the table is the zero skill/negative efficiency combination, where 38% of funds reside.[12] This reveals that the negative average abnormal return in our sample of mutual funds, -1.58%, is driven primarily by funds' inability to fully capture the risk premia associated with the factors to which they are exposed.

4.2 The persistence of skill and efficiency

There is a large literature (e.g., see Brown and Goetzmann, 1995 and Carhart, 1997 for early work in this area) on whether the performance of mutual fund managers is persistent, that is, whether past performance is predictive of future performance. We investigate this in our sample for each of the three measures of ability: skill, efficiency, and abnormal returns. Given our tree model uses the same tree structure for each calendar year, we

[12]Table A6 reports corresponding results when using the random forest or boosted regression tree methods. The results are all similar, except that the proportion of fully efficient funds falls to around one third and the proportion of funds with significant negative efficiency increases to close to two thirds.

study the persistence of yearly aggregates of these measures.[13] We do so via a simple panel autoregressive (AR) model of the form:

$$X_{i,y+1} = c + \phi X_{i,y} + e_{i,y+1} \tag{10}$$

where $X_{i,y}$ denotes one of the measures of ability: skill ($\psi_{i,y}$), aggregate efficiency ($\Delta_{i,y}$) or abnormal return ($\alpha_{i,y}$) in year y.

Table 9 reveals a striking difference in the persistence of these measures of ability. Skill is essentially uncorrelated from year to year, with an AR coefficient of 0.04, and an R^2 of 0.2%. Efficiency, on the other hand, is more persistent, with an AR coefficient of 0.16 and R^2 of 2.8%. Abnormal returns, which are the sum of these two measures, lie in the middle, with an AR coefficient of 0.11 and an R^2 of 1.3%.[14] Thus the component of abnormal returns that drives its persistence from year to year is efficiency.

4.3 Predicting abnormal returns with skill and efficiency

Mutual fund investors cannot consume skill and efficiency separately; what they take to the bank is the sum of these, the abnormal return. We next show that even if interest is purely in abnormal returns, investors can improve their decisions by decomposing abnormal returns into skill and efficiency.

The differences in persistence across the measures of skill documented in the previous section suggests that they may have differing predictive power for future abnormal returns. We investigate this in Table 10 via simple linear regressions.[15] The first column of Table

[13] Doing this at the monthly level introduces look-ahead bias in months other than December.

[14] Results using the random forest and boosted regression tree models, reported in Table A7, reveal the same pattern.

[15] Kaniel *et al.* (2023) consider sophisticated neural networks for forecasting abnormal mutual fund returns, based on past returns and a large number of stock and fund characteristics. This paper complements the work by Kaniel *et al.* in that we find evidence that one component of abnormal returns (efficiency) is more persistent than the other, and thus a sophisticated forecasting framework like theirs may be improved by exploiting this feature.

Table 9: Persistence of measures of mutual fund manager ability.
This table presents parameter estimates from a panel autoregressive model of order one for annualized measures of mutual fund manager ability, given in the column titles, estimated using the tree model described in Section 3.3. All t-statistics use White (1980) standard errors.

	Skill	*Efficiency*	*Abnormal Return*
AR(1)	0.039	0.162	0.110
(t-stat)	(5.614)	(24.300)	(11.765)
Const	1.373	-2.160	-1.021
(t-stat)	(31.909)	(-65.404)	(-27.480)
N	44,031	44,031	44,031
R^2 (%)	0.164	2.763	1.294

10 replicates the AR model for abnormal returns in Table 9 and is included for ease of reference. The second column decomposes lagged abnormal returns into lagged skill and lagged efficiency, and two key findings emerge: firstly, the R^2 from this model increases from 1.3% to 1.9%, a sizable improvement in predictive power. This increase is driven by the model assigning different weights to lagged skill and efficiency: skill has coefficient of 0.09 while the coefficient on efficiency is over double that, at 0.19. A test that these coefficients are equal, and thus that this model collapses to the AR(1) in column 1, rejects the null at the 0.01 level.[16] Thus if abnormal returns are the sum of a persistent component (efficiency) and a transitory component (skill), a better forecast of the sum can be obtained by emphasizing the persistent component and de-emphasizing the transitory component.

Next we investigate whether there is *asymmetry* in the predictive power of these measures of ability. This is motivated by evidence, see Brown and Goetzmann (1995) and Carhart (1997) for example, that persistence in mutual fund manager performance is mostly

[16]Recall that since abnormal returns are the sum of skill and efficiency, we cannot include all three measures of ability in the same regression, as that leads to perfect multicollinearity.

Table 10: Predicting abnormal returns.
This table presents results from panel regressions to predict one-year-ahead abnormal returns. The first specification is a first-order autoregression, and corresponds to specification (1) in Table 9. The second specification uses lagged skill and efficiency as predictors. The third specification decomposes lagged abnormal returns into positive and negative components. The fourth specification decomposes skill and efficiency into their positive and negative components. All t-statistics use White (1980) standard errors.

	(1)	(2)	(3)	(4)
Abnormal Return	0.110			
(t-stat)	11.765			
Skill		0.090		
(t-stat)		9.771		
Efficiency		0.186		
(t-stat)		15.307		
AbRet*(AbRet>0)			0.069	
(t-stat)			3.851	
AbRet*(AbRet≤0)			0.142	
(t-stat)			9.163	
Skill*(Skill>0)				0.157
(t-stat)				11.698
Skill*(Skill≤0)				0.000
(t-stat)				0.015
Eff*(Eff>0)				0.020
(t-stat)				0.781
Eff*(Eff≤0)				0.275
(t-stat)				18.184
Const	-1.021	-0.803	-0.830	-0.763
(t-stat)	-27.480	-18.160	-12.347	-10.506
N	44,031	44,031	44,031	44,031
R^2 (%)	1.294	1.861	1.374	2.384

driven by the continued poor performance of the pooly-performing funds. In column three of Table 10 we decompose lagged abnormal returns into positive and negative components. Consistent with the existing literature, we see that the coefficient on lagged negative abnormal returns is larger than that on lagged positive abnormal returns (0.14 versus 0.07, a difference that is significant at the 0.01 level). It is interesting to note that while decomposing lagged abnormal returns into positive and negative components improves the performance of the model, increasing the R^2 from 1.3% to 1.4%, the decomposition in column 2, into skill and efficiency, leads to a much larger improvement in performance, with an R^2 of 1.9%.

In the last column of Table 10 we decompose lagged skill and efficiency into positive and negative parts. The fit of this model is improved relative to treating negative and positive parts the same (the R^2 rises from 1.9% to 2.4%) and this gain comes from identifying the persistent parts of each of these measures. Table 10 shows that it is positive skill and negative efficiency that are predictive for future abnormal returns. The coefficients on negative skill and positive efficiency are not different from zero (t-statistics of 0.02 and 0.78 respectively). This is strong evidence for combining the known asymmetry in the predictive power of positive and negative past performance with the decomposition of abnormal returns into skill and efficiency proposed in this paper.[17]

4.4 Common and relative (in)efficiency

Finally, we consider disaggregating the efficiency of a given fund into a component that is common to all funds and the "relative efficiency" of the fund when compared with the average fund. Recall that the usual Fama-MacBeth (1973) second-stage regression estimates the *average* risk premia earned by all mutual funds, allowing for an estimate of

[17] Table A8 presents corresponding results using the Forest and Boosted regression tree models, and shows that the conclusions are very similar using these other two methods.

the efficiency for an average fund with fund i's factor exposures:

$$\overline{\Delta}_{i,t} \equiv \boldsymbol{\beta}_{i,t-1}^{\top}(\boldsymbol{\lambda}_t^{OLS} - \boldsymbol{\lambda}_t^{Stocks}) \equiv \boldsymbol{\beta}_{i,t-1}^{\top}\bar{\boldsymbol{\eta}}_t \tag{11}$$

All of the cross-sectional variation in $\overline{\Delta}_{i,t}$ is due to variation in factor exposures.

The regression tree allows for estimates of risk premia to vary across funds by exploiting the cross-sectional variation in the state variables used to form the tree. This allows us to estimate:

$$\boldsymbol{\eta}_{i,t}^{Tree} = \boldsymbol{\lambda}_{i,t}^{Tree} - \boldsymbol{\lambda}_t^{Stocks} \tag{12}$$

We can then decompose the aggregate inefficiency of a fund into a component coming purely from its exposures, and a component measuring the efficiency of fund i relative to the average mutual fund:

$$\begin{aligned} \Delta_{i,t} &= \boldsymbol{\beta}_{i,t-1}^{\top}(\boldsymbol{\lambda}_{i,t}^{Tree} - \boldsymbol{\lambda}_t^{Stocks}) \\ &= \overline{\Delta}_{i,t} + \boldsymbol{\beta}_{i,t-1}^{\top}(\boldsymbol{\lambda}_{i,t}^{Tree} - \boldsymbol{\lambda}_t^{OLS}) \end{aligned} \tag{13}$$

Table 11 shows that average efficiency, $\overline{\Delta}_{i,t}$, is useful for predicting future abnormal returns, and improves the R^2 relative to the AR(1) from 1.3% to 1.5%.[18] Column 3 of Table 11 shows that relative efficiency, $\Delta_{i,t} - \overline{\Delta}_{i,t}$ is also useful, with a t-statistic of 6.9, and increases the R^2 to 1.9%. Thus both the component of efficiency driven purely by factor exposures, and the component that measures how efficient the fund is in comparison with the average fund, help explain future abnormal returns.

[18] The model in column 2 of Table 11 is directly comparable that in column 2 of Table 10, when we note that the fit of a model using skill and efficiency is identical to that using abnormal returns and efficiency, given the collinearity of these three measures of ability. We observe that decomposing abnormal returns into skill and efficiency is more beneficial when using the tree-based model than simple OLS, with the R^2 of the former model being 1.86% and the latter being 1.54%.

Table 11: Decomposing mutual fund efficiency: Average and relative efficiency This table presents results from panel regressions to predict one-year-ahead abnormal returns. The first specification is a first-order autoregression, and corresponds to specification (1) in Table 9. The second specification adds lagged efficiency estimated using the "OLS" method. The third specification further adds the difference between efficiency estimated using a regression tree and estimated using OLS. All t-statistics use White (1980) standard errors.

	(1)	(2)	(3)
Abnormal Return	0.110	0.100	0.090
(t-stat)	(11.765)	(10.587)	(9.747)
Average efficiency		0.085	0.070
(t-stat)		(6.548)	(5.192)
Relative efficiency			0.135
(t-stat)			(7.133)
Const	-1.021	-0.809	-0.877
(t-stat)	(-27.481)	(-15.192)	(-15.883)
N	44,031	44,031	44,031
R^2 (%)	1.294	1.531	1.908

5 Conclusion

The idea that the performance of a fund manager should be gauged by risk-adjusted performance goes back over half a centery (Jensen, 1968) and has been extensively used by investors and academics. The number of risks that are considered in such models has grown (e.g., see Fama and French, 1992; Carhart, 1997; Fama and French, 2015 and others) but the assumption that simply earning the risk premia associated with the risks to which a fund is exposed is not a skill has gone largely unquestioned.

Work on implementation costs (e.g., Keim and Madhavan, 1997; Wermers, 2000; Novy-Marx and Velikov, 2016; Patton and Weller, 2020) has revealed that earning the risk premium associated with certain factors, momentum in particular, is difficult, especially for

large investors. We show that the presence of such costs, the familiar abnormal return, or alpha, is made up of two components: "efficiency," which is the sum of the fund's (in)efficiencies across risk factors, weighted by the fund's exposures to those factors, and "skill," a component that is unrelated to factor exposures.

Using a variety of methods, including ex ante sorts and machine learning tools, we find significant heterogeneity in efficiency in our sample of over 5,000 U.S. mutual funds. Put simply, some funds earn a higher risk premium for exposure to the same risk factor. For example, funds with returns that are well-explained by the Fama and French (1992)-Carhart (1997) model accrue 7.7% per year for exposure to the market factor, while funds with more idiosyncratic returns earn only 4.2% for the same exposure.

We study the time series properties of skill and efficiency, and find that efficiency is much more persistent than skill, consistent with the former being a more fundamental property of the manager, or manager's investment process, and with the latter being more attributable to luck. Consistent with this, we find that abnormal returns can be better predicted using lagged skill and efficiency separately, rather than combined. In fact, the out-of-sample predictive gains from decomposing lagged abnormal returns into skill and efficiency are seven times as large as the gains from exploiting the well-known result (see Brown and Goetzmann, 1995; Carhart, 1997) that abnormal return persistence is driven by poor past performers. Combining these two ideas leads to even greater gains.

Appendix

A Data cleaning methods

Following Berk and van Binsbergen (2015) and Patton and Weller (2020), we clean the CRSP mutual fund data in two stages: first at the individual fund level to address missing and erroneous data, and then at the fund group level to apply higher-level sample filters. For simplicity, we refer to 'fund groups' as 'funds' throughout the main text.

A.1 Fund-level data cleaning

Our data cleaning begins at the individual fund level, addressing inconsistencies and missing values prior to aggregation. We use mutual fund monthly return data from CRSP, covering January 1970 through December 2023, resulting in 9,399,372 observations. We first flag fund-months with returns reported less frequently than monthly. For these, we mark the return as missing if neither adjacent month contains a valid, non-zero return. These infrequent annual returns represent 0.71% of all fund-month observations. We next clean the total net asset (TNA) values, which are crucial for value-weighting fund returns. Following established practices, we treat TNAs less than or equal to $100,000 as bottom-coded and mark them as missing. We also exclude extremely high values exceeding $1 trillion. These filters result in 12.29% of TNA observations being marked as missing. To recover missing TNAs, we use a three-step interpolation process. First, we project missing TNAs by compounding the last known TNA using cumulative fund returns since that point — though this does not capture cash flows. Second, when a subsequent valid TNA exists, we calculate the discrepancy between the predicted and actual TNA as the ratio of actual TNA to predicted TNA (minus one) to infer implied flows, which we assume are evenly distributed over time. Specifically, the predicted TNA is adjusted by multiplying it by

$(1 + discrepancy)^{s/\Delta t}$ where s is the number of months since the last valid TNA and Δt is the total gap length. If no subsequent TNA is available, we assume zero discrepancy. Third, we apply the first two steps in reverse to impute TNAs prior to the first observed value. After interpolation, we again flag any TNAs below \$100,000 or above \$1 trillion as invalid. These procedures reduce missing TNA values to 3.6% of the dataset.

To adjust for share class differences in fees, we convert net returns into gross returns by adding one-twelfth of the annual expense ratio, following Fama and French (2010). Initially, 19.02% of observations in the fund summary file have missing or non-positive expense ratios. We first fill these gaps using the nearest non-missing value within each CRSP fund number group, reducing the missing rate to 10.6%. We then merge the monthly return and summary files by fund number and calendar quarter, assigning expense ratios to 74.88% of fund-month observations. For the remaining unmatched records, we merge again using fund number and calendar year, averaging the expense ratios within each fund-year. This additional merge raises coverage to 85.49% (or 8,035,470 observations). We remove 136 observations with expense ratios over 50%, as these are likely data errors.

Following Patton and Weller (2020), we remove return outliers, specifically any observation with an absolute return exceeding 100%, which is likely due to reporting mistakes. This results in the exclusion of 162 such observations. We also eliminate any fund-month observation ever flagged as an ETF, ETN, or VAU fund during the fund's life, based on the "et_flag" variable. These filters exclude approximately 14.3% of the total observations.

A.2 Aggregating funds into fund groups

To identify funds belonging to the same fund group, we begin by imputing missing fund names using the nearest valid name within each CRSP fund number group. Out of 2,713,579 observations in the fund summary file, this procedure recovers names for 23,184 entries. We remove 285 observations where a name cannot be recovered. A similar approach is used

to fill missing ticker symbols, recovering 136,537 out of 379,100 missing tickers. Unlike missing names, we retain records with missing tickers.

We then follow the two-step grouping algorithm detailed by Patton and Weller (2020), Berk and van Binsbergen (2015), and Pastor *et al.* (2015):

1. Parsing fund names: We identify share classes using three mutually exclusive rules:

   - If the fund name contains a semicolon and the part following the last semicolon does not include a forward slash, we use the portion before the last semicolon as the group name.
   - If the name contains a forward slash and the substring after the last slash lacks spaces or semicolons, we use the part before the slash.
   - If neither condition applies, we assume the name does not include a share class indicator.

   Before applying these rules, we adjust for fund names where slashes are not share class delimiters—e.g., "Franklin/Templeton" or "M/M." To address these, we replace slashes with backslashes in known patterns like T/F, T/E, M/M, L/S, Small/Mid, Long/Short, S/T, and L/T.

2. Creating equivalence classes: We define funds as equivalent if they share an adjusted name or a ticker symbol. Equivalence is treated as transitive: if Fund A shares a name with Fund B, and Fund B shares a ticker with Fund C, all three are grouped together.

Applying this procedure reduces the 72,070 unique fund IDs in the CRSP monthly return file to 28,192 fund groups. Of the total observations in the monthly return file, 5,615 could not be matched to a group and were removed.

A.3 Fund group-level cleaning

We construct fund group returns and TNAs using component fund IDs. Group returns are value-weighted using one-month lagged TNAs. If a group contains only a single fund, we retain its observation even when the lagged TNA is missing. Group-level TNA is the sum of current TNAs across all component funds. These steps result in 2,548,148 monthly fund-group observations.

To address incubation bias, as noted in Fama and French (2010), we retain only observations after a fund group surpasses a TNA threshold of $10 million (in December 2023 dollars). We do not drop post-threshold data if the TNA later falls below this cutoff. Excluding fund groups that never reach $10 million eliminates 2.22% of group-month observations, while dropping pre-threshold months removes an additional 4.76%.

We apply additional filters to exclude non-equity and international funds. Specifically, we drop any fund whose name includes terms such as "ETF," "ETN," "exchange-traded fund," "exchange traded fund," "exchange-traded note," "exchange traded note," "iShares," and "PowerShares" (not case sensitive). We also remove funds with terms suggesting non-domestic equity exposure or fixed-income strategies—for instance: "international," "intl," "bond," "emerging," "frontier," "rate," "fixed income," "commodity," "oil," "gold," "metal," "world," "global," "China," "Europe," "Japan," "real estate," "absolute return," "government," "exchange," "euro," "India," "Israel," "treasury," "Australia," "Asia," "pacific," "money," "cash," "yield," "U.K.," "UK," "kingdom," "municipal," "Ireland," "LIBOR," "govt," "obligation," "mm," "m/m," and "diversified" (but not "diversified equity"), and "short term" (not case sensitive). These exclusions reduce the number of valid fund groups from 14,268 to 6,366, leaving 953,001 non-missing monthly return observations for the full period from December 1970 to December 2023. Next, we restrict our sample to fund groups with at least two years of monthly data during the

1970–2023 period. This filter yields a sample of 5,722 mutual fund groups with 942,779 valid monthly return observations.

To construct realized betas, we require daily returns for mutual fund groups. Using the fund groups and their component funds identified earlier, we retrieve daily returns and TNAs for the component funds from CRSP and compute daily fund group returns as (daily) TNA-weighted averages of the component funds' daily returns. We then restrict the sample to fund groups with at least two years of daily data during the 1999–2023 period. This filter yields a final sample of 5,046 mutual fund groups with valid daily return observations. For consistency, we also retain the same 5,046 fund groups in the monthly sample over 1999–2023, rather than using the full set of 5,722.

B Mutual fund realized betas

The standard method for estimating factor exposures is to use a rolling window of 36 or 60 months of returns, see Ferson (2019) and Kaniel *et al.* (2023) for example. The beta for month t is estimated using data from $t-36$ to $t-1$:

$$R_{i,s} - R_s^f = \alpha_{i,t-1} + \boldsymbol{\beta}_{i,t-1}^\top \mathbf{F}_s + \varepsilon_{i,s}, \quad \text{for } s = t-36, ..., t-1 \tag{14}$$

In the popular Fama and MacBeth (1973) procedure, the estimated factor exposures, $\boldsymbol{\beta}_{i,t-1}$, are then used as regressors in the second-stage, cross-sectional, regressions:

$$R_{i,t} - R_t^f = \lambda_{0,t} + \boldsymbol{\lambda}_t^\top \boldsymbol{\beta}_{i,t-1} + u_{i,t}, \quad \text{for } i = 1, 2, ..., N_t \tag{15}$$

where N_t denotes the number of funds available at time t.

With the availability of daily mutual fund returns, we can exploit advances in the econometrics of high frequency data by estimating factor exposures over shorter windows

of time, obtaining more accurate estimates of these important quantities, see Barndorff-Nielsen and Shephard (2004). Lewellen and Nagel (2006) use this "short window regression" approach for stock portfolios, and Akbas and Genc (2020) use it for U.S. mutual funds. Let d denote days and let D_t denote the number of trade days in months $\{t-2, t-1, t\}$, a number that is typically around 65, then we run the regression:

$$R_{i,d} - R_d^f = \tilde{\alpha}_{i,t-1} + \boldsymbol{\beta}_{i,t-1}^\top \mathbf{F}_d + \varepsilon_{i,d}, \quad \text{for } d = 1, 2, ..., D_{t-1} \tag{16}$$

The slope coefficients in equation (16) are a timely and accurate estimate of the exposures of the fund, and avoid the need to specify a parametric model for their movements over time.

C Simulation-based testing of homogeneity

In Section 3.2 we consider tests of equal risk premia across ex ante groups of mutual funds. For a single risk factor, e.g. MKT, we consider:

$$\mathbb{H}_0 \quad : \quad \lambda_g^{\text{MKT}} = \lambda_h^{\text{MKT}} \text{ for all } g, h \tag{17}$$

$$\text{vs.} \qquad \mathbb{H}_1 \quad : \quad \lambda_g^{\text{MKT}} \neq \lambda_h^{\text{MKT}} \text{ for some } g, h \tag{18}$$

where g and h indicate groups of mutual funds. Given that we use quintile sorts, we test a total of $q = 4$ restrictions. For the joint test, across all parameters, we test a total of $(K+1) \times 4 = 20$ restrictions. Asymptotic p-values for these tests are obtained from the χ_q^2 distribution.

To confirm that the asymptotic critical values lead to tests with satisfactory size control, we construct "placebo" groupings of mutual funds and run the tests on those groups. The placebo groupings are formed in an identical way to those formed in Section 3.2, except

Table A1: Comparing asymptotic and simulation-based tests.
This table considers test of equal parameters across quintile groups of mutual funds. The first five rows test each parameter of the model separately, while the last row tests all parameters jointly. Placebo groupings of mutual funds are formed by randomly assigning funds to groups, and averaging results across 1,000 random groupings. The first column of this table presents the rejection rates for these placebo groupings using χ^2 critical values at the 0.05 nominal level. The next two rows present the asymptotic χ^2 critical values for these tests and the critical values obtained from the distribution of test statistics across the 1,000 placebo groupings. The last five columns report the test statistics using each of the five sorting variables considered in Section 3.2.

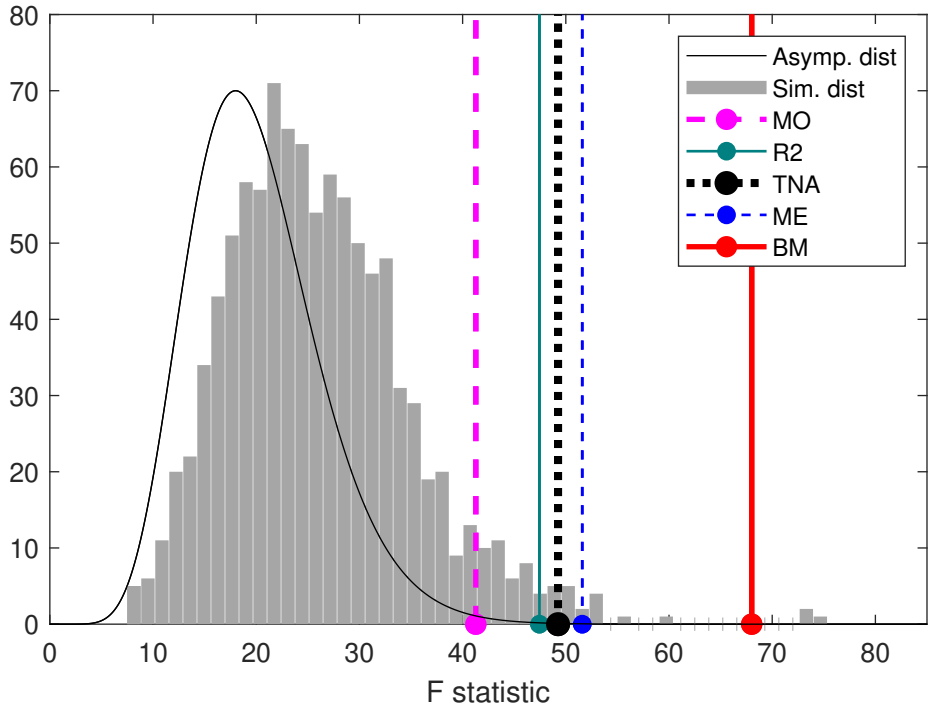
	Placebo rejection rate	Critical value		Test statistic				
		Asymp.	Sim.	TNA	R2	ME	BM	MO
MKT	0.072	9.488	10.297	6.912	26.665	17.310	10.462	11.366
SMB	0.065	9.488	10.286	8.017	8.906	7.683	8.290	3.842
HML	0.059	9.488	9.8092	5.251	4.880	6.192	3.148	2.790
MOM	0.063	9.488	9.9488	10.333	4.296	5.239	9.897	2.073
Const	0.075	9.488	10.418	9.261	2.206	1.091	0.092	2.814
Joint	0.212	31.410	37.897	50.528	50.721	50.989	49.623	34.486

that we use *iid* draws from the $\mathcal{U}(0,1)$ distribution as the sorting variable, with the same missing data structure present in the real sorting variable imposed on the placebo sorting variable. Thus in this analysis all groups are random subsets of the population of funds, and the rejection rate of the test across 1,000 random groupings should approximately equal the nominal size of the test.

The first column of Table A1 shows that the rejection rates for tests of homogeneity for a single variable (the first five rows) are close to the 5% nominal level, ranging from 5.9% to 7.5%, however the joint test across all variables has a rejection rate of 21.2%, far above the nominal level.[19] Figure 4 compares the asymptotic distribution and the simulation-based

[19]The size control of the tests of homogeneity for a single factor are better because they only involve

Figure 4: Distribution of joint tests of homogeneity.
This figure plots the asymptotic distribution of the joint test of homogeneity of all lambdas across quintile groups formed using a sorting variable, and the simulation-based distribution using random group assignments. The F-statistics associated with tests of homogeneity across quintiles formed using one of five sorting variables (MO, R^2_{FF4}, TNA, ME and BM) are presented as vertical lines.



distribution for the joint test, and shows that the latter is a right-shift of the former.

Given the importance of the joint test for determining whether mutual funds significantly differ in their ability to accrue risk premia, and the poor size control of this test when using the asymptotic critical value, we base our conclusions from the analysis in Section 3.2 using the simulation-based test results.

D Regression tree details

To be completed...

estimating a 5×5 asymptotic covariance matrix, while the joint test across all parameters involves a 25×25 asymptotic covariance matrix. The latter is a challenging task with only 300 months of data. The simulation-based critical values capture the difficulty of estimating this matrix.

E Additional tables

Table A2: Differences in earned risk premia, monthly betas.
This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior 36 monthly returns, for comparison with the results using"realized betas" presented in Table 2. The first column uses 269 stock portfolios described in Section 2.1, and represents the "on-paper" risk premia for these factors. The second column presents estimates using our sample of 5,046 mutual funds, also described in Section 2.1. The third column presents the difference between the two estimates, and can be interpreted as the average "(in)efficiency" of mutual funds in our sample for a specific risk factor. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all four differences are zero. All t-statistics and p-values use Newey-West(1987) standard errors with five lags.

	Stocks	**Mutual funds**	**Difference**
MKT	8.758	6.029	-2.730
(t-stat)	(2.722)	(1.944)	(-4.18)
SMB	3.064	3.443	0.379
(t-stat)	(1.676)	(1.988)	(0.576)
HML	1.903	0.545	-1.357
(t-stat)	(0.738)	(0.199)	(-1.593)
MOM	2.340	0.387	-1.954
(t-stat)	(0.820)	(0.118)	(-1.298)
R^2	0.283	0.514	
$\bar{N}$	269	2,245	
Joint p-value			0.000

Table A3: Differences in earned risk premia, no intercept.
This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B The first column uses 269 stock portfolios described in Section 2.1, and represents the "on-paper" risk premia for these factors. The second column presents estimates using our sample of 5,046 mutual funds, also described in Section 2.1. For both samples the intercept is suppressed. The third column presents the difference between the two estimates, and can be interpreted as the average "(in)efficiency" of mutual funds in our sample for a specific risk factor. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all four differences are zero. All t-statistics and p-values use Newey-West(1987) standard errors with five lags.

	Stocks	**Mutual funds**	**Difference**
MKT	8.970	6.945	-2.026
(t-stat)	(2.760)	(2.056)	(-5.209)
SMB	2.554	3.095	0.542
(t-stat)	(1.361)	(1.730)	(0.727)
HML	1.268	-0.695	-1.962
(t-stat)	(0.532)	(-0.264)	(-1.999)
MOM	1.693	0.087	-1.606
(t-stat)	(0.553)	(0.028)	(-1.484)
R^2	0.326	0.542	
$\bar{N}$	269	2,472	
Joint p-value			0.000

Table A4: Heterogeneity in mutual fund efficiency: Sorts by MO of holdings. This table presents estimates of risk premia associated with the four risk factors in the Fama-French (1992)-Carhart (1997) model, as well as abnormal returns in the first row, estimated using the Fama-MacBeth (1973) procedure. Factor loadings ("betas") are estimated using the prior three months of daily returns, as described in Appendix B. The first column uses all mutual funds for which we have holdings data, and differs slightly from the second column in Table 2 due to missing data. The next five columns report parameter estimates separately for each quintile of mutual funds sorted by the value-weighted average of the twelve-month momentum (MO) of their stock holdings. The last column presents the p-value associated with a test that the parameter estimates in that row are equal across all quintiles. The row labeled R^2 reports the average of the cross-sectional R^2s across the 300 months in the sample. The row labeled $\bar{N}$ reports the average number of assets used to estimate the risk premia each period. The last row reports the p-value on the joint test that all quintiles have the same parameters. All t-statistics and p-values use Newey-West(1987) standard errors with five lags.

		MO Quintiles					**p-vals**	
	All funds	**Low**	**2**	**3**	**4**	**High**	**Asymp.**	**Sim.**
MKT	5.526	5.349	4.857	5.586	4.696	4.174	0.428	0.487
(t-stat)	(1.850)	(1.926)	(1.749)	(1.768)	(1.598)	(1.337)		
SMB	3.038	3.388	2.625	3.122	2.261	2.969	0.594	0.624
(t-stat)	(1.694)	(1.800)	(1.396)	(1.939)	(1.211)	(1.609)		
HML	-0.909	-0.239	-0.056	-1.672	-0.819	-0.427	0.722	0.759
(t-stat)	(-0.343)	(-0.089)	(-0.021)	(-0.569)	(-0.327)	(-0.157)		
MOM	0.348	-0.261	-1.888	-0.542	-1.516	-0.042	0.589	0.615
(t-stat)	(0.113)	(-0.089)	(-0.639)	(-0.160)	(-0.545)	(-0.015)		
Const	1.446	1.367	1.706	1.112	2.335	3.062	0.023	0.033
(t-stat)	(1.890)	(1.384)	(1.709)	(1.686)	(2.591)	(3.272)		
R^2	0.542	0.486	0.496	0.638	0.496	0.465		
$\bar{N}$	2,472	494	495	494	495	494		
Joint p-val							0.023	0.167

Table A5: Optimal number of leaves in the regression tree.
This table presents the number of leaves selected by five-fold cross-validation for each of the years in our sample period. The maximum number of leaves considered is ten. Selecting one leaf corresponds to using OLS on the full sample of funds and not splitting into subgroups.

Year	Number of leaves	Year	Number of leaves
1999	4	2012	6
2000	4	2013	7
2001	1	2014	2
2002	1	2015	2
2003	1	2016	8
2004	6	2017	8
2005	5	2018	2
2006	2	2019	9
2007	2	2020	9
2008	5	2021	5
2009	4	2022	7
2010	3	2023	3
2011	3	Average	4.36

Table A6: Skill, efficiency, and abnormal returns: Forest and BRT models. This table presents results on the proportion of mutual funds that have skill, efficiency, and abnormal returns that are significantly different from zero at the 10% level. Estimates of skill and efficiency are based on the random forest and boosted regression tree (BRT) models described in Section 3.3. "Full efficiency" corresponds to an estimate that is not different from zero. The left panel shows a two-way contingency table for skill and efficiency, and the right column shows the results for abnormal returns.

Panel A: Random forest estimates

		Efficiency				
		Negative	**Full**	**Positive**	**Total**	*Abnormal Return*
Skill	**Negative**	0.008	0.012	0.002	0.022	0.202
	Zero	0.456	0.251	0.006	0.713	0.754
	Positive	0.196	0.052	0.017	0.265	0.045
	Total	0.660	0.315	0.025	1	1

Panel B: Boosted regression tree estimates

		Efficiency				
		Negative	**Full**	**Positive**	**Total**	*Abnormal Return*
Skill	**Negative**	0.009	0.010	0.002	0.022	0.202
	Zero	0.459	0.231	0.009	0.699	0.754
	Positive	0.210	0.050	0.019	0.279	0.045
	Total	0.678	0.291	0.030	1	1

Table A7: Persistence of measures of manager ability: Forest and BRT models. This table presents parameter estimates from a panel autoregressive model of order one for annualized measures of mutual fund manager ability, given in the column titles, estimated using the random forest and boosted regression tree models described in Section 3.3. All t-statistics use White (1980) standard errors.

	Random forest		**Boosted regression tree**		
	Skill	*Efficiency*	*Skill*	*Efficiency*	*Abnormal Return*
AR(1)	0.040	0.148	0.035	0.143	0.110
(t-stat)	5.504	21.441	4.770	19.971	11.765
Const	1.640	-2.439	1.694	-2.493	-1.021
(t-stat)	39.013	-78.501	40.416	-78.650	-27.480
N	44,031	44,031	44,031	44,031	44,031
R^2 (%)	0.174	2.343	0.133	2.185	1.294

Table A8: Predicting abnormal returns: Tree, Forest and BRT models.
This table presents results from panel regressions to predict one-year-ahead abnormal returns. The first specification is a first-order autoregression, and corresponds to (1) in Table 10. The second specification decomposes lagged abnormal returns into positive and negative components and corresponds to (3) in Table 10. Specifications (3) to (5) use lagged skill and efficiency as predictors, as measured by the Tree, Forest, and Boosted regression tree (BRT) models described in Section 3.3. Specifications (6) to (8) decomposes skill and efficiency into their positive and negative components. All t-statistics use White (1980) standard errors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Tree	*Forest*	*BRT*	*Tree*	*Forest*	*BRT*
Abnormal return	0.110							
(t-stat)	11.765							
AbRet*(AbRet> 0)		0.069						
(t-stat)		3.851						
AbRet*(AbRet≤ 0)		0.142						
(t-stat)		9.163						
Skill			0.090	0.093	0.091			
(t-stat)			9.771	10.075	9.835			
Efficiency			0.186	0.186	0.193			
(t-stat)			15.307	14.235	14.542			
Skill*(Skill> 0)						0.157	0.166	0.165
(t-stat)						11.698	12.270	12.261
Skill*(Skill≤ 0)						0.000	-0.003	-0.008
(t-stat)						0.015	-0.159	-0.428
Eff*(Eff> 0)						0.020	-0.037	-0.024
(t-stat)						0.781	-1.230	-0.795
Eff*(Eff≤ 0)						0.275	0.293	0.291
(t-stat)						18.184	18.247	17.997
Const	-1.021	-0.830	-0.803	-0.790	-0.763	-0.763	-0.700	-0.732
(t-stat)	-27.480	-12.347	-18.160	-16.432	-15.636	-10.506	-8.627	-9.037
N	44,031	44,031	44,031	44,031	44,031	44,031	44,031	44,031
R^2 (%)	1.294	1.374	1.861	1.677	1.741	2.384	2.321	2.350

Table A9: Decomposing mutual fund efficiency: Average and relative efficiency
This table presents results from panel regressions to predict one-year-ahead abnormal returns. The first specification is a first-order autoregression, and corresponds to specification (1) in Table 9. The second specification adds lagged efficiency estimated using the "OLS" method. The third specification further adds the difference between efficiency estimated using a regression tree and estimated using OLS. All t-statistics use White (1980) standard errors.

	(1)	(2)	(3)	(4)	(5)
Abnormal return	0.110	0.100	0.090	0.091	0.086
(t-stat)	(11.765)	(10.587)	(9.747)	(9.924)	(9.429)
Average efficiency		0.085	0.070	0.071	0.071
(t-stat)		(6.548)	(5.192)	(5.111)	(5.22)
Relative efficiency (Tree)			0.135		
(t-stat)			(7.133)		
Relative efficiency (Forest)				0.172	
(t-stat)				(4.516)	
Relative efficiency (BRT)					0.228
(t-stat)					(6.100)
Const	-1.021	-0.809	-0.877	-0.842	-0.834
(t-stat)	(-27.481)	(-15.192)	(-15.883)	(-15.334)	(-15.483)
N	44,031	44,031	44,031	44,031	44,031
R^2 (%)	1.294	1.531	1.908	1.726	1.859

References

Akbas, F. and Genc, E. (2020). Do mutual fund investors overweight the probability of extreme payoffs in the return distribution? *Journal of Financial and Quantitative Analysis*, 55(1):223–261.

Aleti, S., Bollerslev, T., and Siggaard, M. (2025). Intraday market return predictability culled from the factor zoo. *Management Science*.

Amihud, Y. and Goyenko, R. (2013). Mutual fund's r^2 as predictor of performance. *Review of Financial Studies*, 26(3):667–694.

Ammann, M., Fischer, S., and Weigert, F. (2020). Factor exposure variation and mutual fund performance. *Financial Analysts Journal*, 76(4):101–118.

Banegas, A., Gillen, B., Timmermann, A., and Wermers, R. (2013). The cross section of conditional mutual fund performance in european stock markets. *Journal of Financial economics*, 108(3):699–726.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance*, 65(1):179–216.

Berk, J. B. and van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1):1–20.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2):1046–1089.

Bollen, N. P. and Busse, J. A. (2001). On the timing ability of mutual fund managers. *Journal of Finance*, 56(3):1075–1094.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. CRC Press.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Brown, S. J. and Goetzmann, W. N. (1995). Performance persistence. *Journal of Finance*, 50(2):679–698.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82.

Daniel, K., Grinblatt, M., Titman, S., and Wermers, R. (1997). Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance*, 52(3):1035–1058.

Daniel, K. and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance*, 52(1):1–33.

Fama, E. F. (1991). Efficient capital markets: Ii. *Journal of Finance*, 46(5):1575–1617.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance*, 65(5):1915–1947.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Fan, J., Ke, Z. T., Liao, Y., and Neuhierl, A. (2022). Structural deep learning in conditional asset pricing. *Available at SSRN 4117882*.

Farrell, M. H., Liang, T., and Misra, S. (2025). Deep learning for individual heterogeneity. *arXiv 2010.14694*.

Ferson, W. (2019). *Empirical Asset Pricing: Models and methods*. MIT Press.

Ferson, W. and Wang, J. L. (2021). A panel regression approach to holdings-based fund performance measures. *Review of Asset Pricing Studies*, 11(4):695–734.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R. and Liu, Y. (2022). Luck versus skill in the cross section of mutual fund returns: Reexamining the evidence. *Journal of Finance*, 77(3):1921–1966.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *Journal of Finance*, 23(2):389–416.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kaniel, R., Lin, Z., Pelger, M., and Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1):94–138.

Keim, D. B. and Madhavan, A. (1997). Transactions costs and investment style: An inter-exchange analysis of institutional equity trades. *Journal of Financial Economics*, 46(3):265–292.

Kelly, B. and Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3-4):205–363.

Koijen, R. S. (2014). The cross-section of managerial ability, incentives, and risk preferences. *Journal of Finance*, 69(3):1051–1098.

Lewellen, J. and Nagel, S. (2006). The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics*, 82(2):289–314.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

Newey, W. K. and West, K. D. (1987). A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Novy-Marx, R. and Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147.

Pastor, L., Stambaugh, R. F., and Taylor, L. A. (2015). Scale and skill in active management. *Journal of Financial Economics*, 116(1):23–45.

Patton, A. J. and Weller, B. M. (2020). What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics*, 137:515–549.

Patton, A. J. and Weller, B. M. (2022). Risk price variation: The missing half of empirical asset pricing. *Review of Financial Studies*, 35(11):5127–5184.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442.

Treynor, J. L. (1966). How to rate management investment funds. *Harvard Business Review*, 43.

Wermers, R. (2000). Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance*, 55(4):1655–1695.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.