De facto time-varying indices-based benchmarks for mutual fund returns

Tingting Cheng[1], Cheng Yan[2], Yayi Yan[3]

[1]School of Finance, Nankai University, Tianjin, China

[2]Essex Business School, Essex University, Colchester, UK

[3]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

Correspondence

Cheng Yan, Essex Business School, Essex University, Colchester, CO4 3SQ, UK.

Email: yancheng54@gmail.com

Abstract

We question time-invariant indices as fund benchmarks and propose a regime-switching methodology to identify time-varying de facto benchmarks from a pool of market-based indices, with or without a risk-free asset. To ameliorate the benchmark mismatch issue, we highlight the importance of using time-varying indices-based benchmarks for fund performance evaluation. Our de facto benchmark captures fund styles better than other benchmark choices, substantially improves the identification of significant fund alphas, and provides better out-of-sample forecasts. We uncover several new findings in terms of fund performance evaluation using our de facto benchmarks.

JEL CLASSIFICATION C15, G11, G12, G23

Any analysis of long-term stock price performance invariably grapples with the choice of an appropriate benchmark. The issue is central in studies of stock market efficiency, such as tests of the profitability of trading strategies. Research on the impact of various managerial decisions, such as equity offerings, dividend initiations or omissions, and share repurchase programs, also faces the problem of measuring stock returns in excess of some normal level.

Chan et al. (2009)

1 INTRODUCTION

The techniques for fund performance evaluation can be classified into two approaches: (1) returns-based performance evaluation and (2) portfolio holding-based performance evaluation.[1] Each approach has its own (dis)advantages. Returns-based approaches rely on less information but can be sensitive to the choice of the benchmark portfolio (Chan et al., 2009; Lehmann & Modest, 1987; Roll, 1978).[2] Holding-based approaches allow a more precise construction of a benchmark to address Roll's (1978) criticism,[3] but holding data (if available) are available on a much less frequent basis and hence have limited usefulness. In this article, we not only question

[1] See Ferson (2010) and Wermers (2011) for references and reviews of the earlier literature. See Wermers (in press) for a recent excellent survey paper.

[2] We follow Cremers et al. (2013, p. 6) and define a benchmark as *"a passively managed portfolio with factor exposures similar to the portfolio whose performance we are evaluating."* We acknowledge, however, that there are other definitions of fund benchmarks in the literature.

[3] For instance, Grinblatt and Titman (1993) circumvent Roll's (1978) criticism by proposing a holding-based performance evaluation approach. Daniel et al. (1997) propose benchmarks based on the characteristics of stocks held by the portfolios that are evaluated.

the appropriateness of commonly used fund benchmarks in the literature but also develop and propose a new benchmark identification method that does not require holding data and yields a much more accurate fund benchmark than returns-based benchmarks. Our solution to the benchmark choice question is a flexible regime-switching methodology based on a pool of popular passive Standard & Poor's (S&P) and Russell indices,[4] and we use our proposed time-varying benchmark to investigate the potential influences of the benchmark choice on fund performance evaluation.

We differ from the performance evaluation literature[5] as we are first, to our knowledge, to

[4] We believe that a fund benchmark is better constructed by passive investable indices (i.e., indices-based benchmark) than a combination of arbitrage pricing theory factors (i.e., factors-based benchmark) as we find statistically significant alphas (i.e., unobserved risk compensation) and correlated residuals when we regress popular indices on factors. Berk and van Binsbergen (2015) and Pástor et al. (2015) provide two additional reasons: (1) Fama–French factors do not take into account transaction cost and (2) some of the factors are discovered later than the mutual fund databases. Hence, identifying tradable time-varying indices-based benchmarks is much simpler and meaningful than trying to identify the potentially numerous time-varying factors.

[5] For instance, a large literature focuses on cross-sectionally controlling for the multiple-hypothesis-testing problem (i.e., skill vs. luck; Blake et al., 2013; Blake et al., 2014; Cai et al., 2018; Cheng & Yan, 2017; Fama & French, 2010; Kosowski et al., 2006; Kosowski et al., 2007; Zhang & Yan, 2018), false discovery (Andrikogiannopoulou & Papakonstantinou, 2019; Bajgrowicz & Scaillet, 2012; Bajgrowicz et al., 2015; Barras et al., 2010; Ferson & Chen, 2020; Yan & Cheng, 2019), and time-varying fund alphas and betas (Avramov & Chordia, 2006; Bollen

focus on de facto time-varying indices-based benchmarks for fund returns. The literature (Sensoy, 2009) focuses on either de jure (i.e., self-designated) and/or time-invariant indices-based benchmarks.

We begin by proposing a regime-switching approach to identify a time-varying indices-based benchmark for US equity mutual funds via minimizing the variance of fund alphas from a pool of 17 popular passive S&P and Russell indices (which are defined on size and value/growth dimensions), with or without a risk-free asset. We find a much higher portion of fund benchmark mismatch in our time-varying setting than in the time-invariant setting in Sensoy (2009). To ameliorate the benchmark mismatch issue, we highlight the importance of fund cash holdings (Panageas & Westerfield, 2009; Sensoy, 2009; Simutin, 2014). We evaluate our choice of indices-based benchmark via: (1) the statistical significance of Fama–French (1993) three-factor loadings in explaining funds' monthly benchmark-adjusted returns[6] and (2) the explanatory power of benchmarks on fund excess returns (i.e., average R^2). Intuitively, we find that S&P 500–related indices (i.e., the sum of S&P 500, S&P 500 Value, S&P 500 Growth) are the most popular indices-based benchmarks for mutual funds. Our empirical results also show that the de facto time-varying indices-based benchmark we identify captures the fund styles better than the

& Whaley, 2009; Cai et al., 2018; Cheng et al., 2021; Christopherson et al., 1998; Ferson & Schadt, 1996; Jones & Mo, 2021; Kacperczyk et al., 2014; Mamaysky et al., 2007, 2008; Pástor et al., 2015).

[6] This is similar to the criterion used in Sensoy (2009), and Ferson (2010, p. 211) justifies it as "*benchmark portfolio that has the same regression betas on the risk factors as the fund is an appropriate benchmark.*"

official/self-declared benchmarks as well as the alternative benchmarks Sensoy (2009) identifies, and they partially overlap with the official/self-declared benchmarks.

How do fund benchmark mismatches affect fund performance evaluation? To answer this question, we estimate fund alphas using de facto time-varying indices-based benchmarks and official/self-declared benchmarks, and compare the estimated fund alphas after that. Our de facto time-varying indices-based benchmarks substantially improve the identification of funds with significant alphas. Using our de facto time-varying indices-based benchmarks instead of the commonly used benchmarks in the literature, we identify a larger portion of statistically significant mutual fund alphas with a smaller magnitude on average. We find higher alpha persistence than that found by traditional indices-based benchmarks and factor-based benchmarks in out-of-sample forecasting. Replacing the factor-based benchmarks with our indices-based benchmarks in the bootstrap approach, we find that "luck" (i.e., sample variability) can explain the positive, but not negative, alphas of funds.

We contribute to the literature in several ways. First, we propose a new fund benchmark (i.e., our de facto time-varying indices-based benchmark) for performance evaluation. The paper most closely related to ours is Sensoy (2009), who asks whether a time-invariant passive mainstream stock index exists that captures the fund characteristics better than the de jure benchmarks (official, or self-designated benchmarks). In parallel, Cremers and Petajisto (2009), Cremers and Pareek (2016), and Cremers et al. (2022) introduce time variations into the holding-based fund benchmark literature.

Second, even though the problem of mismatched indices-based benchmarks has been described in the literature (i.e., Elton et al., 2003; Sensoy, 2009), we show that because the extent of mismatching is so large, even the "corrected benchmarks" in the literature (e.g., Sensoy, 2009) designed to match the fund's exposure to size and value/growth factors achieve only modest success, as long as they are not de facto and time varying like ours.[7]

Finally, we contribute to performance evaluation studies in general, given the importance of benchmark choice in these studies (e.g., Ferson, 2010; Wermers, 2011). By constructing a more accurate benchmark, we achieve the goal of Hunter et al. (2014) in terms of improving fund alpha estimation from a different perspective.

This article fits into several major strands of the mutual fund literature: factors-based benchmark (Carhart, 1997; Fama & French, 1993; Jensen, 1968), indices-based benchmark (Berk & van Binsbergen, 2015; Cremers et al., 2013; Pástor et al., 2015), fund benchmark mismatch (Brown & Goetzmann, 1997; Cooper et al., 2005; Daniel et al., 1997; Elton et al., 2003),[8] and fund performance evaluation (Barras et al., 2010; Cai et al., 2018; Cheng & Yan, 2017; Fama & French,

[7] Our sample covers more than 5000 US mutual funds, which is much larger than the sample of 71 funds in Beber et al. (2021), 108 funds in Elton et al. (2003), 199 funds in Chan et al. (2009), and 1981 funds in Sensoy (2009). It is also more relevant and updated. Although the US Securities and Exchange Commission (SEC) only began requiring mutual funds to report a passive benchmark index in 1998, Chan et al.'s (2009) sample ends in 2001 and Sensoy's (2009) in July 2004.

[8] A key issue in mutual funds research is that funds may engage in window dressing and misclassify themselves (Brown & Goetzmann, 1997). Cooper et al. (2005) note that some mutual funds change their names to reflect a hot investment style and attract more fund flows, which affects the benchmarks investors use to evaluate them.

2010; Ferson & Chen, 2020; Harvey & Liu, 2018; Kosowski et al., 2006; Pástor et al., 2015; Yan & Cheng, 2019).

2 BENCHMARKING METHODOLOGY

Since 1998, US mutual funds have been required to specify a benchmark index against which to compare each fund's performance, in the fund prospectus and other public disclosures. A natural performance measure for investors is simply computing the relative performance of an individual fund to its self-declared benchmark index, which matches the fund's actual investment style. However, Sensoy (2009) finds that about 31% of actively managed mutual funds have a benchmark mismatch, that investors react strongly to a fund's performance relative to its self-declared benchmark index, and that benchmark mismatches are likely driven by the incentive to improve cash flows. Hence, we must first find an appropriate benchmark before evaluating an individual fund's performance.

2.1 Regime-switching approach

In this subsection, we introduce a new approach—regime switching—to identify the latent actual benchmark indices in a time-varying setting. Allowing the benchmark to vary over time is important as empirical evidence suggests fund's style drift and time variation in fund risk taking (Cao et al., 2017; Huang et al., 2011). Before introducing our approach to identify time-varying indices-based benchmarks, we recognize that a good benchmark should explain a higher fraction of the variance of the fund returns and decrease the standard error of the estimate of the fund's abnormal performance (i.e., fund alpha), according to the literature (e.g., Daniel et al., 1997). Such a conceptually simple criterion also motivates the following econometric procedures.

Without loss of generality, we motivate our econometric method by writing:

$$r_{it} - r_{bmk,k,t} = \alpha_i + \epsilon_{it}, \tag{1}$$

where r_{it} is the return of fund i at time t, $r_{bmk,k,t}$ denotes the return of benchmark k at time t, α_i is fund alpha, and ϵ_{it} is a tracking error.[9] We assume that there are K benchmarks in total.

In Equation (1), the indices-based benchmark k for fund i is usually unknown because of the possibility of benchmark mismatch. Sensoy (2009) specifies a time-invariant benchmark index, which produces the highest correlation with fund returns among several benchmarks, as the fund's benchmark. Indeed, Sensoy's (2009) procedure is equivalent to selecting a benchmark index that minimizes the variance of ϵ_{it} in Equation (1). We differ from Sensoy (2009) as in Equation (1) the benchmark index for fund i could be different at each time point.

Let $r_t^{bmk} = [r_{bmk,1,t}, \cdots, r_{bmk,K,t}]'$, and ι_{s_t} $(s_t \in \{1, \cdots, K\})$ be the s_tth column of identity matrix I_K . Then, Model (1) becomes a regime-switching model:

$$r_{it} - \iota_{s_t}' r_t^{bmk} = \alpha_i + \epsilon_{it}. \tag{2}$$

In Equation (2), ι_{s_t} is more like a time-varying indicator function rather than a time-varying coefficient. For example, if the benchmark index for fund i is 1 at time t, we have $\iota_{s_t} = [1, 0, \dots, 0]'$ and $r_{it} - r_{bmk,1,t} = \alpha_i + \epsilon_{it}$. By treating the fund's benchmark indicator s_t as a random variable, our method allows for the latent benchmarks to vary over time. Hence, determining the true indices-based benchmark k is equivalent to estimating the latent variable s_t at time t. We then propose the following estimation procedures based on the maximum likelihood approach, which yields the posterior distribution of benchmark indicator s_t. By the nature of the maximum likelihood estimation method, our proposed procedures automatically determine the

[9] Following Petajisto (2013), we call ϵ_{it} a tracking error. Intuitively, it measures the return of the fund that is not explained by alpha and the fund's benchmark index.

most likely benchmark for each mutual fund and explain highest fraction of the variance of the fund returns (i.e., smallest variance of fund alpha).

Let $\Omega_t = (r'_{bmk,t}, \cdots, r'_{bmk,1}, r_{it}, \cdots, r_{i1})'$ be the vector containing observed data up to time t, and $\theta = (\alpha_i, \sigma_i, P)$ be the set of model parameters, where P is the transition matrix in which the element p_{ij} denotes the probability $p(s_t = j | s_{t-1} = l)$ modeling the dynamic behavior of benchmark switching and σ_i is the standard deviation of ϵ_{it}.[10] The log-likelihood function for the observed data is constructed as

$$\ell(r_{i1}, \dots, r_{iT}) = \log f(r_{i1}|\theta) + \sum_{t=2}^{T} \log f(r_{it}|\Omega_{t-1}; \theta). \qquad (3)$$

The computation details of $\ell(r_{i1}, \dots, r_{iT})$ are presented in Online Appendix A.1.

Consequently, the maximum likelihood estimator $\hat{\theta}$ can be obtained by

$$\hat{\theta} = \arg\max \ \ell(r_{i1}, \dots, r_{iT}).$$

There are $K^2 + 2$ unknown parameters in Model (2). When K is not large, the likelihood function can be maximized through numerical methods, such as the Newton–Raphson method. However, in our real-data sample, $K = 17$, which is relatively large, the usual Newton–Raphson algorithm is not a good choice because the optimization frequently fails for a relatively large number of indices-based benchmarks. Therefore, we design an expectation maximization (EM) algorithm

[10] We can relax the assumption of constant transition probabilities by jointly modeling these probabilities. However, this is not necessary because our goal is to estimate the latent actual benchmarks rather than model the dynamic behavior of a fund's style drifts. In addition, we can relax the time-invariant assumption of σ_i using an autoregressive conditional heteroskedasticity approach.

to estimate the unknown parameters. The details of the EM algorithm are outlined in Online Appendix A.2.

A by-product of parameter estimation is the posterior distribution of latent variable s_t. Using a Markov switching filter and the algorithm developed by Kim (1994), we get three kinds of inferences for benchmark indicator s_t: prediction probability $p(s_t|\Omega_{t-1};\theta)$, filter probability $p(s_t|\Omega_t;\theta)$, and smoothed probability $p(s_t|\Omega_T;\theta)$. Hence, we propose the following three detecting criteria for benchmark selection, which determine the most likely benchmark for each mutual fund:

$$\begin{aligned}\hat{k}_{p,t} &\equiv \arg\max_j \; p(s_t = j|\Omega_{t-1};\theta),\\ \hat{k}_{f,t} &\equiv \arg\max_j \; p(s_t = j|\Omega_t;\theta),\\ \hat{k}_{s,t} &\equiv \arg\max_j \; p(s_t = j|\Omega_T;\theta).\end{aligned} \tag{4}$$

These are all good indicators to determine the true time-varying indices-based benchmark according to our simulation studies provided in Online Appendix B. In the next section, based on the previous detecting criteria, we empirically identify the de facto time-varying indices-based benchmarks for US equity mutual funds.

2.2 Accounting for cash holdings

Before proceeding, we propose the following approach to address the issue of fund cash holdings, the importance of which is mentioned in studies such as Panageas and Westerfield (2009), Sensoy (2009), and Simutin (2014). As shown by Huang et al. (2011), the average proportion of cash holding is 6.26% with a standard deviation 9.04% based on a sample of US actively managed equity mutual funds between 1980 and 2006. Therefore, it is essential to take it into account when identifying the latent indices-based benchmarks. To this end, we modify Model (1) as follows:

$$r_{it} - rf_t = \alpha_i + \omega_i\left(r_{bmk,k,t} - rf_t\right) + \epsilon_{it}, \tag{5}$$

where rf_t is risk-free rate and ω_i denotes the proportion of noncash holding. For example, if $\omega_i =$ 0.9, it means that this fund holds 10% cash and 90% stocks. We call the identified benchmarks the "cash-adjusted benchmarks." The identification procedures for benchmarks $r_{bmk,k,t}$ follows directly from Model (1). For convenience, we refer to the nonadjusted benchmarks $r_{bmk,k,t}$ identified by Model (1) as the "no-cash benchmarks."[11]

3 IDENTIFYING DE FACTO TIME-VARYING INDICES-BASED BENCHMARKS

In this section, we apply our regime-switching estimator to real data to empirically identify the latent indices-based benchmarks. We further follow Sensoy (2009) to compare our identified indices-based benchmarks with the commonly used indices-based benchmarks in the literature. Specifically, we use our proposed regime-switching approach to identify the latent time-varying indices-based benchmarks for US equity mutual funds from a pool of 17 popular passive S&P and Russell indices (defined on size and value/growth dimensions). These 17 indices-based benchmarks, denoted by $k = 1, \cdots, 17$, are S&P 500, S&P 500 Value, S&P 500 Growth, S&P 400, S&P 600, Russell 1000, Russell 1000 Value, Russell 1000 Growth, Russell 2000, Russell 2000 Value, Russell 2000 Growth, Russell 3000, Russell 3000 Value, Russell 3000 Growth, Russell Midcap, Russell Midcap Value, and Russell Midcap Growth. Before we present the empirical results, we outline the data sources and describe the main characteristics of mutual funds in our sample.

[11] The next section shows that cash-adjusted benchmarks capture the fund risk exposure much better.

3.1 Data and descriptive statistics

We obtain the net returns[12] of active US equity mutual funds from the Center for Research in Security Prices (CRSP) Survivor-Bias-Free Mutual Fund database between November 1998 and December 2014.[13] We exclude index funds. To mitigate omission bias (Elton et al., 2001) and incubation and back-fill bias (Evans, 2010), we exclude observations before the reported year when the mutual funds first entered into the database, and the funds that do not report the year of organization. We include only funds that have initial total net assets (TNA) above $10 million and more than 80% of their holdings in equity markets. To avoid a look-ahead bias, we do not exclude funds whose TNA subsequently falls below $10 million. These screens leave us with a sample of

[12] Although our methodology is flexible enough for the gross alphas (Pástor et al., 2015) and value added (Berk & van Binsbergen, 2015), we follow the mainstream literature (e.g., Barras et al., 2010; Ferson & Chen, 2020) and use the net alphas to illustrate our idea. We have no intention to be involved in the re-heated debate on which measure is the right/better measure of fund skills.

[13] We start our sample in 1998, the year the SEC began requiring each fund to report a benchmark index in its prospectus. Cremers and Petajisto (2009, p. 3340) note:

> *Since 1998, the SEC has required each fund to report a benchmark index in its prospectus; however, this information is not part of any publicly available mutual fund database, and prior to 1998, it does not exist for all funds. These self-declared benchmarks might even lead to a bias: some funds could intentionally pick a misleading benchmark to increase their chances of beating the benchmark by a large margin, as discussed in Sensoy (2009).*

5592 mutual funds with at least 18 months of returns data between November 1998 and December 2014. Table 1 presents summary statistics of the mutual funds, which share similar characteristics with the data used in Ferson and Chen (2020) and Cheng et al. (2021). The main characteristics are:

- The median value of average mutual fund returns is slightly positive at 0.2% per month. The range of average returns across mutual funds is $-0.077 \sim 0.062$.
- The median of estimated alpha from the Fama–French three factors for mutual funds is slightly negative (i.e., -0.001).
- The sample volatility of the median mutual fund return is 5.1% per month. Between the 10% and 90% quantiles, the volatility range is 3.1% $\sim$ 7.6% (4.2% $\sim$ 7.0% in Ferson & Chen, 2020) for mutual funds.
- The median autocorrelation for mutual funds is 0.125. The range of autocorrelation for mutual funds is $-0.667 \sim 0.580$.

The data for self-declared indices-based benchmarks and active-share benchmarks are taken from Petajisto (2013), covering an unbalanced panel of 2740 mutual funds. To facilitate the comparison of our estimated indices-based benchmarks with the self-declared indices-based benchmarks and active-share benchmarks, we take the funds' intersection of both data sets, which leaves us with 1647 funds left between November 1998 and September 2009. In addition, we convert quarterly data of active-share benchmarks into monthly data by setting the benchmark index to be same within a quarter. Finally, we follow Cremers and Petajisto (2009) and Sensoy (2009) to obtain monthly returns on S&P and Russell indices from the websites of their parent companies.

3.2 Time-varying popularity of indices-based benchmarks

We begin by computing the percentages of these estimated benchmarks at each period for all 5592 mutual funds. Figures 1 and 2 display the results for no-cash benchmarks and cash-adjusted benchmarks, respectively. From both figures, we can see that the percentages of these 17 indices-based benchmarks fluctuate much over time, which suggests time-varying fund styles for the fund industry. Table 2 presents the time-series averages of the percentages of these 17 indices-based benchmarks. We can see that S&P 500–related indices (i.e., the sum of S&P 500, S&P 500 Value, and S&P 500 Growth) are the most popular indices-based benchmarks for mutual funds.

In addition, we compare our estimated indices-based benchmarks with the cross-sectional self-declared indices-based benchmarks from Cremers and Petajisto (2009) by computing the overlapping ratio (OR) defined as

$$OR_t = n_t^*/N^*,$$

where N^* denotes the number of funds in both data sets (i.e., 1695) and n_t^* denotes the number of funds with the same indices-based benchmark in the two data sets at time t. Figure 3 displays the results are displayed. We find that our identified benchmarks partially overlap with the official/self-declared benchmarks for the same funds, and the ORs of the self-declared indices-based benchmark with our estimated no-cash benchmarks and cash-adjusted benchmarks are similar. Furthermore, the ratios are around 18% before the global financial crisis and about 15% afterward. This indicates that there is a much higher portion of fund benchmark mismatch ($1 -$ OR) in our time-varying setting than in the time-invariant setting in Sensoy (2009), which is about one-third. The time-varying OR also reveals that fund-style drift is more frequent during the financial crisis.

3.3 Covariance comparisons with alternative indices-based benchmarks

In this subsection, we follow Sensoy (2009) and use the Fama–French three-factor regression to explain funds' monthly benchmark-adjusted returns:[14]

$$r_{it} - r_{bmk,k,t} = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \epsilon_{it}. \qquad (6)$$

where r_{it} is fund i's return at time t and $r_{bmk,k,t}$ is the return of indices-based benchmark k at time t; MKT_t, SMB_t, and HML_t denote the Fama–French three factors, which are the market excess return (*MKT*) factor, the small-minus-big (*SMB*) size factor, and the high-minus-low (*HML*) value factor at time t, respectively. The factor loadings in each regression identify differences between the fund's and its benchmark's average exposures to the factors. Based on Model (1), a good benchmark should capture all systematic risk exposure for fund returns. We conjecture that our estimated time-varying indices-based benchmark will induce the smallest significance ratio of the Fama–French three factors.

For comparison, we consider the following set of comparing benchmarks, which includes the risk-free rate, market factor,[15] S&P 500, Sensoy (2009) time-invariant benchmarks, and cash-holding-adjusted, time-invariant benchmarks.[16] Panel A of Table 3 reports statistics from the distribution of regression coefficients and assesses how frequently and on which factors the

[14] A similar empirical model specification is used in Angelidis et al. (2013) and related studies.

[15] Here the market factor is the market excess return factor in the Fama–French three-factor model.

[16] The cash-holding-adjusted, time-invariant benchmarks are obtained based on Sensoy's (2009) time-invariant benchmarks and Equation (5).

differences are significant for our full sample (i.e., 5592 mutual funds) using our estimated no-cash benchmarks and cash-adjusted benchmarks as well as the previously described set of comparing benchmarks. A substantial fraction of funds has significantly different risk exposure on these three factors, even for our estimated no-cash benchmark. Overall, our regime-switching-based benchmarks, especially the cash-adjusted benchmarks, yield the least characteristic difference between fund returns and their corresponding benchmarks among these indices-based benchmarks.

To be specific, a substantial fraction of funds have significantly different market exposure compared to the corresponding indices-based benchmarks.[17] Most differences are negative (except when using the risk-free rate). This pattern is fairly uniform across indices-based benchmarks even for our estimated no-cash benchmark. Thus, we conjecture this is at least partially due to fund cash holdings (Panageas & Westerfield, 2009; Sensoy, 2009; Simutin, 2014).

Fewer funds display significant loadings on *SMB* relative to their benchmarks.[18] Specifically,

[17] The market exposures are 83.10%, 68.10%, 66.11%, 66.61%, 6.96%, 57.47%, and 8.58% for the risk-free rate, market factor, S&P 500, Sensoy (2009) time-invariant benchmarks, cash-holding- adjusted time-invariant benchmarks, no-cash benchmarks, and cash-adjusted benchmarks, respectively. Our regime-switching-based benchmarks achieve the smallest exposure, especially the cash-adjusted benchmarks.

[18] The loadings are 43.40%, 43.40%, 50.29%, 35.25%, 22.34%, 14.72%, and 16.01% for the risk-free rate, market factor, S&P 500, Sensoy (2009) time-invariant benchmarks, cash-holding-adjusted benchmarks, no-cash benchmarks, and cash-adjusted benchmarks, respectively. Our two regime-switching-based benchmarks achieve the smallest loading.

4.04% (3.18%) of funds have positive and significant *SMB* loadings relative to our estimated no-cash benchmark (cash-adjusted benchmark) and 10.68% (12.82%) have negative loadings, and there is more positive risk exposure on *SMB* of funds relative to the comparing benchmarks. The portion of funds with statistically significant *HML* loadings relative to their benchmarks of risk-free rate, market factor, and S&P 500 is about 45%, which decreases to about 30% when we use the Sensoy (2009) time-invariant benchmarks or our estimated no-cash benchmarks, and further drops to about 20% using our estimated cash-adjusted benchmarks.

Overall, the fraction of funds having significantly different risk exposure on the Fama–French three factors is relatively high using these five indices-based benchmarks (risk-free rate, market factor, S&P 500, Sensoy time-invariant benchmarks, and no-cash benchmarks); this is at least partially due to cash holdings. After taking cash holdings into consideration, we can see that the cash-holding-adjusted time-invariant benchmark and the cash-adjusted benchmarks capture the risk exposure on the market factor much better. Furthermore, using our proposed cash-adjusted benchmarks, the exposure on the Fama–French three factors is significant only for a very small fraction of these funds.

Panel B of Table 3 reports performance evaluation results by including the self-declared benchmark and the active-share benchmark as additional comparing benchmarks. For comparison purposes, we consider only the 1647 funds that are in both data sets. The results are qualitatively similar to those in Panel A, and our estimated indices-based benchmarks perform much better in capturing the exposure on the Fama–French three factors than self-declared benchmarks and active-share benchmarks.

4 PERFORMANCE EVALUATION WITH OUR IDENTIFIED BENCHMARKS

In this section, we use our de facto time-varying indices-based benchmarks to evaluate fund performance for both in-sample and out-of-sample analyses.

4.1 Econometric analysis

In this subsection, we outline the econometric advantages of using our identified time-varying benchmarks, in addition to traditional risk factors and time-invariant indices-based benchmarks. Consider the case in which the commonly used benchmark indices have nonzero alphas (Cremers et al., 2013) and the asset pricing errors ϵ_{it} are often correlated across mutual funds. To illustrate the usefulness of our method, we propose the following theoretical framework based on the preceding phenomena and the baseline Model (1).

We model the return of a fund benchmark index by

$$\begin{aligned} r_{bmk,k,t} - rf_t &= x_t'\beta_k + f_t'\lambda_k \\ &= \mathrm{E}(f_t'\lambda_k) + x_t'\beta_k + \nu_{k,t}, \end{aligned} \tag{7}$$

where x_t denotes the pricing factors (e.g., Fama–French factors), β_k is the benchmark-specific factor loadings, f_t denotes unobserved risk factors, λ_k is the corresponding factor loadings, and $\nu_{k,t} = f_t'\lambda_k - \mathrm{E}(f_t'\lambda_k)$. Though the factor structure is prevalent in the recent econometrics literature to capture cross-section dependence (see, e.g., Bai, 2009), we believe that we are the first to introduce it to model the return of a fund benchmark. In the absence of some unknown factors, if we regress benchmark returns $r_{bmk,k,t}$ on observed factors x_t, the estimated intercept is close to $\mathrm{E}(f_t'\lambda_k)$ and may be statistically significantly different from zero. In addition, the benchmark residuals may be correlated due to sharing the common factors f_t. In Online Appendix C.1, we provide empirical evidence that the estimated $\mathrm{E}(f_t'\lambda_k)$ for 17 benchmarks, termed as unobserved risk compensation, is usually positive and sometimes statistically significantly different from zero. In addition, the results show substantial across-benchmark commonality in idiosyncratic errors of benchmarks. In all, introducing unobserved factors helps explain the empirical phenomenon that passive indices are mispriced by traditional factor models. More important, in what follows we demonstrate how these unobserved factors affect the performance evaluation procedures based on traditional factor models.

Substituting $r_{bmk,k,t}$ into Model (1), we obtain a linear factors-based benchmark:

$$\begin{aligned} r_{it} - rf_t &= \alpha_i + x_t'\beta_k + f_t'\lambda_k + \epsilon_{it} \\ &= \alpha_i + E(f_t'\lambda_k) + x_t'\beta_k + \nu_{k,t} + \epsilon_{it}, \end{aligned} \tag{8}$$

which is popular in fund performance evaluation (e.g., capital asset pricing model [CAPM] from Jensen, 1968; Fama–French three-factor model from Fama & French, 1993; Fama–French–Carhart four-factor model [FFC4] from Carhart, 1997). If unobserved factors f_t exist, the ordinary least squares (OLS) estimator of alpha is biased by $\mathrm{E}(f_t'\lambda_k)$. Given the evidence in Online Appendix C.1, if people regress fund returns only on pricing factors, they may overestimate fund alphas because the unobserved risk factors are not taken into account properly. In addition to the biasness of alpha estimator, the existence of unobserved factors f_t results in unexplained covariation among mutual fund residuals. As discussed by Hunter et al. (2014), the correlated residuals from commonly used models reduce the power of such models to separate skilled from unskilled fund managers. Online Appendix C.2 shows how to overcome this difficulty using our proposed methods. In summary, we find that the standard pricing factors leave a significant degree of unexplained covariation among 17 benchmark returns. Furthermore, such unexplained covariation helps explain fund residuals, which motivates our proposed de facto time-varying indices-based benchmarks.

4.2 Alpha estimation diagnostics

In this subsection, we empirically investigate the influence of different indices-based benchmarks on alpha estimation.

4.2.1 Empirical models

To demonstrate the effect of using different benchmarks, we use data on US equity-oriented mutual funds and measure performance using the standard four-factor model (FFC4; Carhart, 1997) as well as models augmented with different benchmarks. For each fund, we use the FFC4 model as our baseline performance evaluation model, against which we test our alternative specification that augments the model with five indices-based benchmarks, which are S&P 500, Sensoy (2009) time-invariant benchmarks, cash-holding adjusted time-invariant benchmarks, no-cash benchmarks, and cash-adjusted benchmarks.

The original FFC4 model (Carhart, 1997) applied to fund i is:

$$r_{it} - rf_t = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + u_i UMD_t + \epsilon_{it}, \qquad (9)$$

where UMD_t denotes the momentum factor monthly return and all variables are previously defined. To demonstrate the advantage of our benchmarking methodology, we consider a modified FFC4 model using unobserved risk-compensation-adjusted benchmark returns:

$$r_{it} - rf_t - \left(r_{bmk,k,t} - rf_t - \mathrm{E}(f_t'\lambda_k)\right) = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + u_i UMD_t + \epsilon_{it}. (10)$$

Equation (10) helps control for unobserved commonalities (arising from sharing unobserved common factors f_t) in fund residuals as shown in Online Appendix C.2. This is done by subtracting unobserved commonalities in matched benchmark indices from fund returns and observing that the other part of $r_{bmk,k,t} - rf_t - \mathrm{E}(f_t'\lambda_k)$ is explained by FFC4 factors (i.e, the mean of $\nu_{k,t}$ is zero). Hence, Model (10) yields the same magnitudes but smaller standard errors of estimated alphas as Model (9).

In addition, we apply a modified FFC4 model using index benchmark-adjusted returns (Angelidis et al., 2013; Sensoy, 2009):

$$r_{it} - r_{bmk,k,t} = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + u_i UMD_t + \epsilon_{it}. \qquad (11)$$

As shown in Online Appendix C.1, if people regress fund returns only on pricing factors, they may overestimate fund alphas because the unobserved risk factors are not taken into account properly. In such a case, the benchmark-adjusted evaluation Model (11) removes unobserved risk compensation $\mathrm{E}(f_t'\lambda_k)$, as well as unobserved commonalities, and thus provides an unbiased alpha estimator $\hat{\alpha}_i$ (both magnitudes and standard errors of estimated alphas differ from Model (9)). As Model (11) is equivalent to Model (9) augmented with a passive factor (i.e., the difference between benchmark return and risk-free rate), according to Pástor and Stambaugh (2012), the passive factor can take into account the time-varying commonalities across funds. In other words, we improve the estimation of alphas from a performance evaluation regression by including the passive factor to reduce the idiosyncratic noise in common.

The alpha estimate in Model (9) can be simply obtained by OLS. For Model (10), we first regress $r_{bmk,k,t}$ on observed pricing factors to get the estimate of unobserved risk compensation $\mathrm{E}(f_t'\lambda_k)$, denoted by $\mathrm{E}\widehat{(f_t'\lambda_k)}$. Then, we replace $\mathrm{E}(f_t'\lambda_k)$ by $\mathrm{E}\widehat{(f_t'\lambda_k)}$ in Model (10) to get the OLS estimate of alpha. For Model (11), we regress benchmark-adjusted returns $r_{it} - r_{bmk,k,t}$ on observed pricing factors to get a bias-corrected estimate of alpha.

4.2.2 Results

Table 4 presents the results of the alpha estimation. The third column (α_{rf}) reports the percentage of funds having statistically significant alphas using Model (9), and the fifth and seventh columns (α_1^{bmk} and α_2^{bmk}, respectively) report the percentage of funds having significant alphas using Models (10) and (11). To compute these percentages, we count the number of significant *p*-values, which are those below 2.5%, and divide by the total number of funds within a particular category. Furthermore, within each group (e.g., no-cash benchmarks), the first (fourth) row reports the percentage of funds with significant positive (negative) alphas and the average values of those alphas, and the second (third) row reports the percentage of funds with insignificant positive (negative) alphas and the corresponding average values of those alphas.

Before comparing the effects of different benchmarks, we compare benchmark augmented models with the standard four-factor (Carhart, 1997) model. Comparing the third and fifth columns in Table 4, we find that Model (10) with all five indices-based benchmarks identifies a larger portion of significant alphas than Model (9). For example, Model (9) indicates that 1.93% (22.59%) of funds have the ability to generate significant positive (negative) alphas, whereas Model (10) with the five indices-based benchmarks indicates that 2.38% (24.73%), 2.93% (25.72%), 4.20% (22.94%), 4.10% (32.31%), and 6.13% (32.85%) of funds generate significant positive (negative) alphas, respectively. The larger portion of significant alphas identified by Model (10) results because the standard errors of estimated alphas are different due to the reduced idiosyncratic disturbances captured by good benchmarks, although the magnitudes of alphas for Model (9) (α_{rf}) and Model (10) (α_1^{bmk}) are equal. As we show in Online Appendix C.1, unobserved risk compensation is usually positive, and thus the alphas estimated from the FFC4 model overestimate the magnitude of alphas. The improved identification of significant alphas using our time-varying indices-based benchmarks indicates that we should remove this part in estimated fund alphas.

Moreover, the improved identification using our two time-varying indices-based benchmarks is much higher than those using the Sensoy (2009) time-invariant benchmark, cash-holding-adjusted time-invariant benchmark, and S&P 500 index. This result suggests that our estimated benchmarks can better filter out the unobserved common shocks to fund returns and thus increase the power of *t*-statistics by removing the volatility of idiosyncratic errors for those funds.

Looking at the average magnitude of the significant positive and negative alphas with each benchmark, we find that using our time-varying indices-based benchmarks results in a magnitude that is relatively smaller compared with those by the Sensoy (2009) time-invariant benchmark, cash-holding-adjusted time-invariant benchmark, and S&P 500 index. For example, Model (10) with our two benchmarks shows that the average of the significant positive (negative) alphas is 0.0085 (−0.0060) and 0.0083 (−0.0056), whereas Model (10) with the other three benchmarks shows the average is 0.0121 (−0.0072), 0.0106 (−0.0068), and 0.0093 (−0.0068), respectively.

In addition, our two time-varying indices-based benchmarks achieve a much higher average adjusted R^2 than the other two benchmarks as well as the standard FFC4 model, although it is well known that the explanatory power of FFC4 is very high (see, e.g., Hunter et al., 2014). For example, the average adjusted R^2 for Model (11) with our two benchmarks ranges between 61.90% and 63.26%, whereas it is only 53.19%, 53.50%, 55.91%, and 54.47% for the S&P 500, Sensoy (2009) time-invariant benchmark, cash-holding-adjusted time-invariant benchmark, and FFC4 model, respectively.

To check the robustness of these results, we consider the three models based on the Fama–French three-factor and Fama–French five-factor models. Tables 5 and 6 present the results, respectively. The results are qualitatively similar to those Table 4. Within a smaller sample, we also check alternative benchmarks such as the self-declared benchmark and the active-share benchmark and find qualitatively similar results to those in Table 4. These results are available upon request.

In summary, our de facto benchmark captures fund styles better than other benchmark choices and substantially improves the identification of significant fund alphas. Specifically, we identify a larger portion of statistically significant mutual fund alphas with a smaller magnitude, on average.

4.3 Alpha persistence in out-of-sample forecasting

In this subsection, we follow Hunter et al. (2014) and design a simple out-of-sample experiment to test whether our de facto time-varying indices-based benchmarks can generate higher alpha persistence than other benchmarks in out-of-sample forecasting. As there are missing common factors in the standard linear factor models (as shown in Online Appendix C), spurious alphas emerge. As spurious alpha fluctuates over time, the estimated alphas from mismatched benchmarks are less persistent. Hence, we conjecture that our de facto time-varying indices-based benchmarks can generate higher alpha persistence than other benchmarks in the literature.

Specifically, we follow Hunter et al. (2014) and conduct a simple out-of-sample test of performance persistence using Models (9) and (11) as follows:

- Step 1: We estimate fund alphas between January 1999 and December 2002 using Model (11) using six benchmarks (i.e., risk-free rate, S&P500, Sensoy time-invariant benchmarks, cash-holding-adjusted time-invariant benchmarks, no-cash benchmarks, and cash-adjusted benchmarks).[19]
- Step 2: We rank all US equity mutual funds into quartiles using the alpha *t*-statistic from each model for each benchmark.[20]

[19] We conduct an out-of-sample forecasting experiment using self-declared benchmarks and active-share benchmarks, though the out-of-sample period is much shorter due data availability. The untabulated results suggest higher alpha persistence using our identified time-varying indices-based benchmarks than these two benchmarks in out-of-sample forecasting.

[20] In untabulated results, we find qualitatively similar results using estimated alpha instead of alpha *t*-statistics, which is consistent with Kosowski et al. (2006) and Hunter et al. (2014).

- Step 3: For funds in each quartile, we use Models (9) and (11), respectively, to estimate fund alphas and then compute the percentages of significant positive and negative alphas, which are used to measure their performance in the next nonoverlapping 4 years (i.e., January 2003–December 2006).
- Step 4: We repeat Steps 1–3 in a nonoverlapping rolling fashion until we obtain fund alphas over all out-of-sample years.

To examine the out-of-sample forecasting performance using different benchmarks, we define performance discrepancy as the difference between the percentages of significant positive (or negative) alphas in the first quartile and fourth quartile. Specifically, we compute two measures: (1) percentage of signifanct positive alphas in the first quartile minus those in the fourth quartile and (2) percentage of significant negative alphas in the first quartile minus those in the fourth quartile. The larger the magnitude of performance discrepancy, the better the corresponding benchmark performs.

Table 7 presents the results from this exercise using Model (9). The table shows that, in general, funds in the first quartile exhibit much more (less) significant positive (negative) alphas than funds in the fourth quartile. Moreover, the performance discrepancies between funds in the first and fourth quartiles ranked by our de facto time-varying indices-based benchmarks are larger than those ranked by other indices-based benchmarks. For instance, in 2003–2006, out-of-sample forecasting performance discrepancy ranked by our cash-adjusted time-varying benchmarks is 17.16% for positive alphas (−34.48% for negative alphas), whereas out-of-sample forecasting performance discrepancy ranked by S&P 500 is 7.18% for positive alphas (−20.96% for negative alphas).

Overall, all the evidence shows that there is higher alpha persistence using our de facto time-varying indices-based benchmarks than traditional indices-based benchmarks and factor-based benchmarks in out-of-sample forecasting. Hence, our de facto time-varying indices-based benchmarks can be used to categorize funds into different subgroups or gauge the optimal number of fund subgroups (Yan & Cheng, 2019), for instance.

4.4 Skills versus "luck" or unobserved risk compensations

As shown in the previous section, using our de facto benchmark, we can identify a larger portion of statistically significant mutual fund alphas with a smaller average magnitude. The next question is: Are these significant (and positive) alphas due to genuine managerial skills or pure sampling variability (i.e., luck) or unobserved risk compensations? In the literature, the unobserved risk compensations are ignored and researchers usually employ the traditional cross-sectional bootstrap approach combined with a factors-based benchmark to distinguish skilled alphas from lucky alphas (e.g., Cai et al., 2018; Fama & French, 2010; Kosowski et al., 2006). However, as we mention earlier, if a fund has risk exposure on missing factors, the traditional factors-based benchmark treats pricing factors as random errors and thus leaves risk compensation unexplained, which then becomes abnormal returns (i.e., alphas) for funds. Although the literature argues that the bootstrap approach combined with the FFC4 model has a great ability to distinguish skilled alphas from lucky alphas, our analysis suggests that the alphas surviving in the bootstrap test could be due to unobserved risk compensation rather than genuine managerial skills.

We use the intrafund bootstrap scheme[21] with the FFC4 model and a new CAPM model based

[21] Our results are robust to some other bootstrap procedures, including the interfund bootstrap and pooled bootstrap considered by Cai et al. (2018). The results are available upon

on our identified time-varying indices-based benchmarks:

$$r_{it} - rf_t = \alpha_i + \beta_i(r_{bmk,k,t} - rf_t) + \epsilon_{it}. \tag{12}$$

Before presenting the results, we give a brief review of the intrafund bootstrap scheme as follows:

- Step 1: Estimate Model (9) for the ith funds, $i = 1,\cdot\cdot\cdot,N$, and obtain coefficient estimates $\{\hat{\alpha}_i, \hat{\beta}_i, \hat{s}_i, \hat{h}_i, \hat{u}_i\}$, residuals $\{\hat{\epsilon}_{it}\}$. Then, sorting the estimated fund alphas $\hat{\alpha}_i$ and the associated t-statistics $\hat{t}_{\hat{\alpha}_i}$, we can obtain the 1% to 99% quantiles accordingly. The procedure based on Model (12) is the same as for Model (9).
- Step 2: For the ith fund, $i = 1,\cdot\cdot\cdot,N$, generate the bootstrap residuals $\{\epsilon_{it}^b\}_{t=1}^T$ from the empirical distribution of residuals $\{\hat{\epsilon}_{it}\}_{t=1}^T$, where b is the bootstrap index. Then, generate a time series of pseudo-monthly excess returns for this fund under the null hypothesis (i.e., $\alpha_i = 0$):

  $$r_{it}^b = \hat{\beta}_i\, MKT_t + \hat{s}_i SMB_t + \hat{h}_i HML_t + \hat{u}_i UMD_t + \epsilon_{it}^b.$$

  We then reestimate the model based on r_{it}^b and obtain N simulated alphas.
- Step 3: Repeat Step 2 for B (= 1000) times. We then obtain the distribution of these cross-sectional draws of alphas and their t-statistics.
- Step 4: Compute the quantiles of the cross-sectional alphas and t-statistics with the simulated samples.

Thus, Steps 2 and 3 generate artificial funds where the alphas are zero across funds and periods. Funds with positive and negative alphas exist in the bootstrap samples but are due to pure sampling variability (i.e., luck). Step 4 compares the distribution of estimated alphas and t-

request.

statistics with their bootstrap counterparts under the null hypothesis of zero performance, which allows us to make an inference of abnormal returns.

Tables 8 and 9 present the results based on estimated alphas and their *t*-statistics, respectively. In Table 8, the first two columns report the selected quantiles and the cumulative distribution function of the actual estimated alphas at selected quantiles when they are ranked from highest to lowest, and the following two columns report the cumulative distribution function of the simulated luck distribution as well as the *p*-values that correspond to the selected quantiles of the distribution of the simulated alphas based on Model (9). The remaining six columns report the results based on CAPM with our no-cash benchmarks and cash-adjusted benchmarks, respectively.

According to Table 8, several interesting observations can be made. First, the tails of the mutual fund cross-sectional alpha distribution include relatively large values, as the bottom and top funds have alpha estimates of −4.39% and 4.12% per month, respectively, based on the FFC4 model, whereas the median fund in our sample has an alpha of −0.13% per month. Second, for the above-median funds, the results based on our proposed benchmarks are different from those based on the FFC4 model. Using the factors-based benchmarks, the performance of the above-median mutual funds is not subject to the critique of sampling variability (i.e., luck), as we can reject the null hypothesis that the performance of the majority of the top 20% mutual funds is an artifact of sampling variability, albeit with a few exceptions. Using our proposed benchmarks to control for unobserved risk compensation, we cannot reject the null hypothesis that the performance of the above-median mutual funds is an artifact of sampling variability. The reason for the difference is that, as evidenced in Online Appendix C, the traditional factors-based benchmark suffers from the missing-factor problem and might give misleading conclusions. Third, for the below-median funds, both factors-based benchmarks and our proposed benchmarks reach the same conclusion that the significant and negative alphas are due to inferior managerial skills, at the conventional level of significance for most of the mutual funds using our bootstrap schemes. Moreover, the explanatory power of time-varying indices-based benchmarks on fund excess returns is much higher than that of the FFC4 model, which implies that our identified time-varying benchmarks are more appropriate for fund performance evaluation.

Note that estimated alphas $\widehat{\alpha}$ measure only the economic size of abnormal performance but suffer from a potential lack of precision in the construction of confidence intervals, whereas their *t*-statistics $\hat{t}_{\widehat{\alpha}}$ are a pivotal statistic with better sampling properties (Kosowski et al., 2006). Table 9 presents results for funds ranked by their *t*-statistics of alphas. Compared with Table 8, we find that with factors-based benchmarks, only the top five funds exhibit significant positive alphas instead of the top 20% funds discovered using estimated alphas. Using the benchmarks identified by us, the results are the same no matter whether we use alphas or *t*-statistics. This again shows the robustness and advantages of our proposed benchmarks.[22]

To sum, our results differ from the literature that argues that a minority of funds have superior skilled alphas (Kosowski et al., 2006). Using our de facto benchmarks, we demonstrate that the positive (even significant) abnormal returns for funds might be unobserved risk compensation. Therefore, the key take-away is that we should choose an appropriate benchmark before evaluating fund performance.

5 CONCLUDING REMARKS

As the variance of fund alpha increases whenever the indices-based benchmark is mismatched (Daniel et al., 1997), we propose a regime-switching approach to identify a time-varying indices-based benchmark by minimizing the variance of fund alphas, using a pool of 17 popular passive S&P and Russell indices. We evaluate the choice of indices-based benchmarks via: (1) the statistical significance of Fama–French three-factor loadings in explaining funds' monthly benchmark-adjusted returns and (2) the explanatory power of benchmarks on fund excess returns (i.e., average R^2).

Intuitively, we recognize that S&P 500–related indices (i.e., the sum of S&P 500, S&P 500 Value, and S&P 500 Growth) are the most popular indices-based benchmarks for mutual funds. We find a much higher portion of fund benchmark mismatch in our time-varying setting than in

[22] To check the robustness of these results, we replace Model (9) with Fama–French three-factor and five-factor models, and find results qualitatively similar to those in Tables 8 and 9. Our results, available upon request, are also robustness to different lengths of data records, which includes funds that have at 18, 30, and 60 months of observations.

the time-invariant setting in Sensoy (2009). We also highlight the importance of fund cash holdings (Panageas & Westerfield, 2009; Sensoy, 2009; Simutin, 2014). Our empirical results show that our identified de facto time-varying indices-based benchmarks capture fund styles better than the official/self-declared benchmarks as well as the alternative benchmarks identified by Sensoy (2009). However, we noted that our benchmarks partially overlap with the official/self-declared benchmarks.

We demonstrate that our de facto time-varying indices-based benchmarks significantly improve the identification of funds with positive and negative alphas. Using our identified benchmarks instead of the commonly used benchmarks in the literature, we identify a larger portion of statistically significant mutual fund alphas with a smaller magnitude on average. Previous studies may significantly overestimate fund alphas, and fund investors should take caution. We find higher alpha persistence using our de facto time-varying indices-based benchmarks than traditional indices-based benchmarks in out-of-sample forecasting. Replacing the factor-based benchmarks with our indices-based benchmarks in the bootstrap approach, we find that luck (i.e., sample variability) can explain the positive alphas, but not the negative alphas, of mutual funds.

Overall, the evidence documented here underscores the importance of benchmark choice in performance evaluation studies (e.g., Ferson, 2010; Wermers, 2011). Unlike ex ante self-declared benchmarks, the benchmarks identified by us and other researchers (e.g., Sensoy, 2009) are ex post, which is a possible caveat. Although we focus on mutual funds, our regime-switching model can be used for other types of funds (especially funds without holding data) as well. Because some mutual funds in the Morningstar database have two designated benchmarks, a fruitful direction for future research is to investigate whether a mix of several benchmarks instead of a single benchmark better captures the style of some funds, and whether it can further improve the identification of fund alphas. Because we consider a pool of only 17 popular passive S&P and Russell indices defined based on size and value/growth dimensions (Sensoy, 2009), other possible research directions are to gauge the optimal number of passive indices (a smaller benchmark set may lead to more reliable results) and to include alternative indices defined based on other factor-based dimensions (e.g., momentum and quality investing).

ACKNOWLEDGMENTS

We are grateful to grants from Zhejiang Natural Science Foundation of China (ZJNSFC Grant LZ20G010002), National Natural Science Foundation of China (NNSFC Grant 72173068,71803091), the Ministry of Education in China Project of Humanities and Social Sciences (Grant 8YJC790015), Social Science Development Fund of Nankai University (Grant ZB22BZ0321), and the Fundamental Research Funds for the Central Universities (Grant ZB22000101). We alone are responsible for all errors.

REFERENCES

Andrikogiannopoulou, A., & Papakonstantinou, F. (2019). Reassessing false discoveries in mutual fund performance: Skill, Luck, or lack of power? *Journal of Finance*, *74*(5), 2667–2688.

Angelidis, T., Giamouridis, D., & Tessaromatis, N. (2013). Revisiting mutual fund performance evaluation. *Journal of Banking & Finance*, *37*(5), 1759–1776.

Avramov, D., & Chordia, T. (2006). Asset pricing models and financial market anomalies. *Review of Financial Studies*, *19*(3), 1001–1040.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279.

Bajgrowicz, P., & Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, *106*(3), 473–491.

Bajgrowicz, P., Scaillet, O., & Treccani, A. (2015). Jumps in high-frequency data: Spurious detections, dynamics, and news. *Management Science*, *62*(8), 2198–2217.

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

Beber, A., Brandt, M. W. , Cen, J., & Kavajecz, K. A. (2021). Mutual fund performance: Using bespoke benchmarks to disentangle mandates, constraints and skill. *Journal of Empirical Finance*, *60*, 74–93.

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

Blake, D., Caulfield, T., Ioannidis, C., & Tonks, I. (2014). Improved inference in the evaluation of mutual fund performance using panel bootstrap methods. *Journal of Econometrics*, *183*(2), 202–210.

Blake, D., Rossi, A. G., Timmermann, A., Tonks, I., & Wermers, R. (2013). Decentralized investment management: Evidence from the pension fund industry. *Journal of Finance*, *68*(3), 1133–1178.

Bollen, N. P., & Whaley, R. E. (2009). Hedge fund risk dynamics: Implications for performance appraisal. *Journal of Finance*, *64*(2), 985–1035.

Brown, S. J., & Goetzmann, W. N. (1997). Mutual fund styles. *Journal of Financial Economics*, *43*(3), 373–399.

Cai, B., Cheng, T., & Yan, C. (2018). Time-varying skills (versus luck) in U.S. active mutual funds and hedge funds. *Journal of Empirical Finance*, *49*(12), 81–106.

Cao, C., Iliev, P., & Velthuis, R. (2017). Style drift: Evidence from small-cap mutual funds. *Journal of Banking & Finance, 78*, 42–57.

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

Chan, K. C., Dimmock, S. G., & Lakonishok, J. (2009). Benchmarking money manager performance: Issues and evidence. *Review of Financial Studies*, *22*(11), 4553–4599.

Cheng, T., & Yan, C. (2017). Evaluating the size of the bootstrap method for fund performance evaluation. *Economics Letters*, *156*(C), 36–41.

Cheng, T., Yan, C., & Yan, Y. (2021). Improved inference for fund alphas using high-dimensional cross-sectional tests. *Journal of Empirical Finance*, *61*(3), 57–81.

Christopherson, J. A., Ferson, W. E., & Glassman, D. A. (1998). Conditioning manager alphas on economic information: Another look at the persistence of performance. *Review of Financial Studies*, *11*(1), 111–142.

Cooper, M. J., Gulen, H., & Rau, P. R. (2005). Changing names with style: Mutual fund name changes and their effects on fund flows. *Journal of Finance*, *60*(6), 2825–2858.

Cremers, M., Fulkerson, J. A., & Riley, T. B. (2022). Benchmark discrepancies and mutual fund performance evaluation. *Journal of Financial and Quantitative Analysis*, *57*(2), 543–571.

Cremers, M., & Pareek, A. (2016). Patient capital outperformance: The investment skill of high active share managers who trade infrequently. *Journal of Financial Economics*, *122*(2), 288–306.

Cremers, M., & Petajisto, A. (2009). How active is your fund manager? A new measure that predicts performance. *Review of Financial Studies*, *22*(9), 3329–3365.

Cremers, M., Petajisto, A., & Zitzewitz, E. (2013). Should benchmark indices have alpha? Revisiting Performance Evaluation. *Critical Finance Review*, *2*(1), 1–48.

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

Elton, E. J., Gruber, M. J., & Blake, C. R. (2001). A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar Mutual Fund databases. *Journal of Finance*, *56*(6), 2415–2430.

Elton, E. J., Gruber, M. J., & Blake, C. R. (2003). Incentive fees and mutual funds. *Journal of Finance, 58*(2), 779–804.

Evans, R. B. (2010). Mutual fund incubation. *Journal of Finance, 65*(4), 1581–1611.

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

Ferson, W. E. (2010). Investment performance evaluation. *Annual Review Financial Economics*, *2*(1), 207–234.

Ferson, W. E., & Chen, Y. (2020). How many good and bad fund managers are there, really? In C. Lee (Ed.), *Handbook of financial econometrics, mathematics, statistics, and technology* (pp. 3753–3827) World Scientific Publishing.

Ferson, W. E., & Schadt, R. W. (1996). Measuring fund strategy and performance in changing economic conditions. *Journal of Finance*, *51*(2), 425–461.

Grinblatt, M., & Titman, S. (1993). Performance measurement without benchmarks: An examination of mutual fund returns. *Journal of Business*, *66*(1), 47–68.

Harvey, C. R., & Liu, Y. (2018). Detecting repeatable performance. *Review of Financial Studies*, *31*(7), 2499–2552.

Huang, J., Sialm, C., & Zhang, H. (2011). Risk shifting and mutual fund performance. *Review of Financial Studies*, *24*(8), 2575–2616.

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

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

Jones, C. S., & Mo, H. (2021). Out-of-sample performance of mutual fund predictors. *Review of Financial Studies*, *34*(1), 149–193.

Kacperczyk, M., Van Nieuwerburgh, S., & Veldkamp, L. (2014). Time-varying fund manager skill. *Journal of Finance*, *69*(4), 1455–1484.

Kim, C. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, *60*(1-2), 1–22.

Kosowski, R., Naik, N. Y., & Teo, M. (2007). Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics*, *84*(1), 229–264.

Kosowski, R., Timmermann, A., Wermers, R., & White, H. (2006). Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance*, *61*(6), 2551–2595.

Lehmann, B. N., & Modest, D. M. (1987). Mutual fund performance evaluation: A comparison of benchmarks and benchmark comparisons. *Journal of Finance*, *42*(2), 233–265.

Mamaysky, H., Spiegel, M., & Zhang, H. (2007). Improved forecasting of mutual fund alphas and betas. *Review of Finance*, *11*(3), 359–400.

Mamaysky, H., Spiegel, M., & Zhang, H. (2008). Estimating the dynamics of mutual fund alphas and betas. *Review of Financial Studies*, 21(1), 233–264.

Panageas, S., & Westerfield, M. M. (2009). High-water marks: High risk appetites? Convex compensation, long horizons, and portfolio choice," *Journal of Finance*, *64*(1), 1–36.

Pástor, L., & Stambaugh, R. F. (2012). On the size of the active management industry. *Journal of Political Economy*, *120*(4), 740–781.

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

Petajisto, A. (2013). Active share and mutual fund performance. *Financial Analysts Journal*, *69*(4), 73–93.

Roll, R. (1978). Ambiguity when performance is measured by the securities market line. *Journal of Finance*, *33*(4), 1051–1069.

Sensoy, B. A. (2009). Performance evaluation and self-designated benchmark indexes in the mutual fund industry. *Journal of Financial Economics*, *92*(1), 25–39.

Simutin, M. (2014). Cash holdings and mutual fund performance. *Review of Finance*, *18*(4), 1425–1464.

Wermers, R. (2011). Performance measurement of mutual funds, hedge funds, and institutional accounts. *Annual Review Financial Economics*, *3*(1), 537–574.

Wermers, R. (in press). Active investing and the efficiency of security markets. *Journal of Investment Management*.

Yan, C., & Cheng, T. (2019). In search of the optimal number of fund subgroups. *Journal of*

Empirical Finance, *50*(1), 78–92.

Zhang, H., & Yan, C. (2018). A skeptical appraisal of the bootstrap approach in fund performance evaluation. *Financial Markets, Institutions and Instruments*, *27*(2), 49–86.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

FIGURE 1 Time-varying percentages of estimated no-cash benchmarks. We use our proposed regime-switching approach with no-cash benchmarks to identify the latent time-varying indices-based benchmarks for 5592 mutual funds. This figure plots the percentage of the 17 benchmarks at each date in our sample

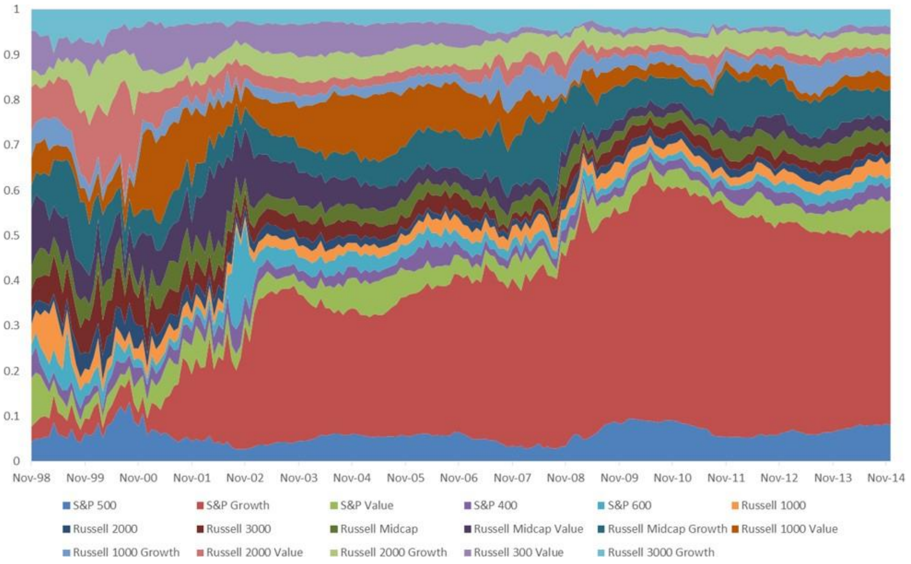

FIGURE 2 Time-varying percentages of estimated cash-adjusted benchmarks. We use our proposed regime-switching approach with cash-adjusted benchmarks to identify the latent time-varying indices-based benchmarks for 5592 mutual funds. This figure plots the percentage of the 17 benchmarks at each date in our sample

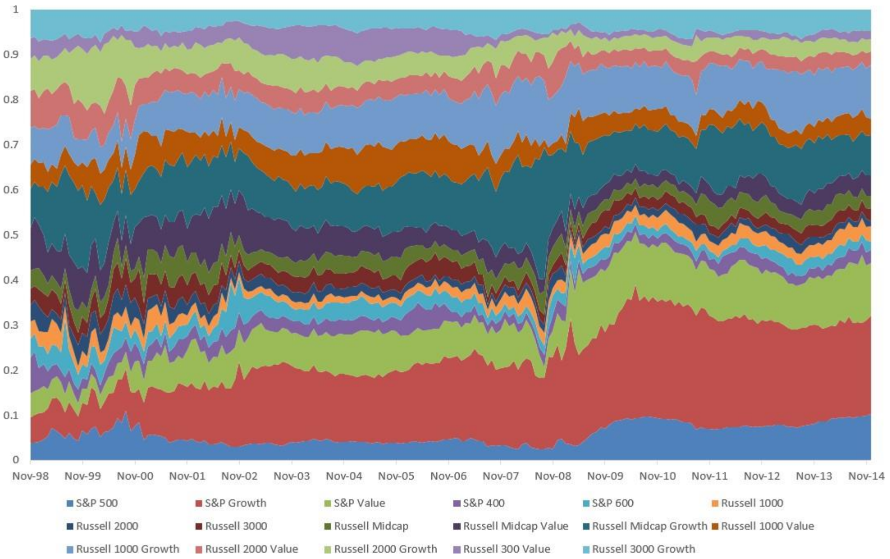

FIGURE 3 Overlapping ratio of estimated benchmarks and self-declared benchmarks. We compare our estimated benchmarks with the self-declared benchmarks of 1695 funds at each date in our sample. In the figure, the solid line denotes the overlapping ratio of our estimated no-cash benchmarks with the self-declared benchmarks. The dashed line denotes the overlapping ratio of our estimated cash-adjusted benchmarks with the self-declared benchmarks

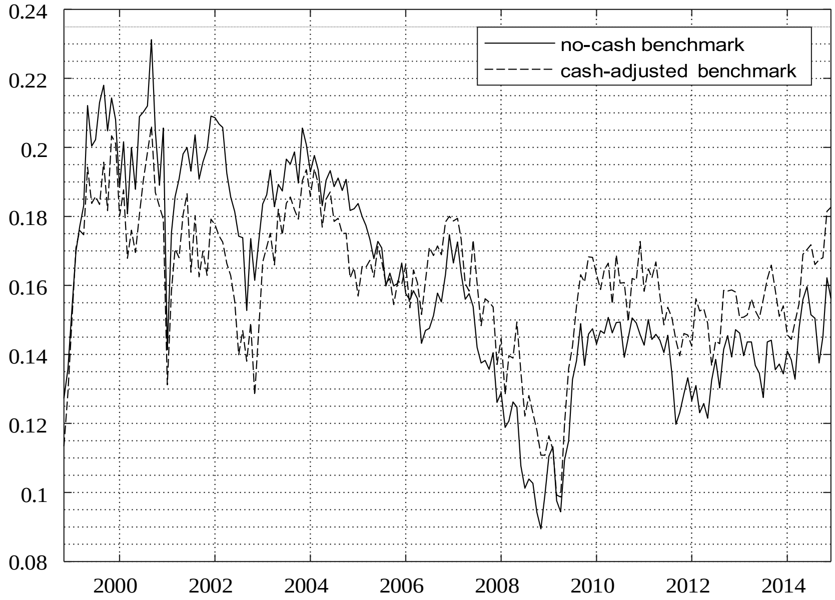

TABLE 1 Summary statistics

	Mutual funds (min. 18 obs.)				
Quantile	**Obs.**	**Mean**	***SD***	**Rho1**	$\hat{\alpha}_{ols}$
Top	194	0.062	0.242	0.580	0.042
1%	194	0.017	0.132	0.415	0.014
5%	194	0.012	0.091	0.324	0.008
10%	194	0.009	0.076	0.263	0.005
20%	182	0.006	0.065	0.204	0.003
30%	146	0.005	0.059	0.171	0.001
Median	94	0.002	0.051	0.125	-0.001
30%	57	-0.001	0.045	0.070	-0.003
20%	45	-0.004	0.040	0.020	-0.004
10%	31	-0.007	0.031	-0.073	-0.006
5%	24	-0.011	0.024	-0.149	-0.009
1%	20	-0.025	0.012	-0.409	-0.016
Bottom	18	-0.077	0.005	-0.667	-0.044

Note: This table reports monthly returns for mutual funds over November 1998–December 2014, measured in excess of the 1-month return of a 3-month Treasury bill. The values at the cutoff points for various quantiles of the cross-sectional distributions of the sample of funds are reported. Each column is sorted on the statistic shown. Obs. is the number of available monthly returns, where a minimum of 18 observations are required. Mean is the sample mean return, *SD* is the sample standard deviation of return, and Rho1 is the first-order sample autocorrelation. The alpha estimates are based on ordinary least squares (OLS) regressions using the Fama–French three factors for mutual funds.

TABLE 2 Time-varying popularity of indices-based benchmarks

Benchmark	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
No cash	5.93	4.37	31.86	2.66	2.91	2.84	6.75	3.23	2.12	3.98	4.70	3.67	4.14	3.95	3.57	5.35	7.97
Cash adjusted	5.56	8.65	17.25	2.92	3.08	2.45	5.71	9.78	2.20	4.82	5.37	3.18	3.33	4.63	3.55	5.68	11.86

Note: We use our proposed regime-switching approach with no-cash benchmarks and cash-adjusted benchmarks to identify the latent time-varying indices-based benchmarks for 5592 mutual funds. This table presents the time-series averages of percentage (%) of estimated indices-based benchmarks for mutual funds in our sample.

TABLE 3 Covariance differences between funds and alternative benchmarks

Benchmark	$\boldsymbol{\beta_i}$			$\boldsymbol{s_i}$			$\boldsymbol{h_i}$		
	$\% \neq \mathbf{0}$	$\% > \mathbf{0}$	$\% < \mathbf{0}$	$\% \neq \mathbf{0}$	$\% > \mathbf{0}$	$\% < \mathbf{0}$	$\% \neq \mathbf{0}$	$\% > \mathbf{0}$	$\% < \mathbf{0}$
Panel A: Full sample									
Risk-free rate	83.10	82.49	0.61	43.40	23.98	19.42	46.82	19.53	27.29
MKT	68.10	11.39	56.71	43.40	23.98	19.42	46.82	19.53	27.29
S&P 500	66.11	14.45	51.66	50.29	47.53	2.75	45.92	17.61	28.31
Sensoy (2009)	66.61	9.01	57.60	35.25	10.46	24.79	31.49	10.98	20.51
Modified Sensoy (2009)	6.96	3.90	3.06	22.34	8.23	14.11	26.14	8.74	17.40
No cash	57.47	7.51	49.96	14.72	4.04	10.68	30.90	5.42	25.48
Cash adjusted	8.58	0.68	7.90	16.01	3.18	12.82	18.96	3.36	15.59
Panel B: 1647 funds only									
Active-share	59.48	8.72	50.76	36.12	19.11	17.02	41.89	15.43	26.46
Self-declared	60.99	8.80	52.20	41.10	21.49	19.61	50.25	16.65	33.60
Risk-free rate	85.15	85.08	0.07	53.64	30.93	22.71	51.48	20.04	31.43
MKT	62.51	11.03	51.48	53.64	30.93	22.71	51.48	20.04	31.43
S&P 500	59.55	13.77	45.78	49.89	47.66	2.24	52.20	18.24	33.96
Sensoy (2009)	62.22	5.98	56.24	27.25	8.72	18.53	30.57	11.18	19.39
Modified Sensoy (2009)	10.08	0.31	9.77	20.05	3.79	16.80	24.92	3.42	21.50
No cash	49.24	5.19	44.05	14.06	6.92	7.14	29.20	5.41	23.79
Cash adjusted	16.08	5.34	10.74	15.93	4.97	10.96	22.06	4.40	17.66

Note: This table reports covariance differences between funds and alternative benchmarks for the full sample in Panel A and the 1647 funds with self-declared benchmarks and active-share benchmarks in Panel B. Factor loadings come from fund-by-fund Fama–French three-factor regressions to explain monthly benchmark-adjusted returns: $r_{it} - r_{bmk,k,t} = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \epsilon_{it}$, where r_{it} is fund i's return at time t and $r_{bmk,k,t}$ is the return of indices-based benchmark k at time t, and MKT_t, SMB_t, and HML_t denote the Fama–French three factors, which are the market excess return (*MKT*) factor, the small-minus-big (*SMB*) size factor, and the high-minus-low (*HML*) value factor at time t, respectively. For each benchmark, the columns display the percentage of funds for which the factor loading is significantly different from, greater than, and less than zero. Sensoy (2009) denotes the benchmark obtained by Sensoy (2009). Modified Sensoy (2009) denotes the cash-holding- adjusted time-invariant benchmark.

TABLE 4 Percentage of funds with (in)significant estimated alpha using Fama–French–Carhart four factors

	Model (9)		**Model (10)**		**Model (11)**	
Benchmark	$\boldsymbol{\alpha_{rf}}$	$\mathbf{Av}(\boldsymbol{\alpha_{rf}})$	$\boldsymbol{\alpha_1^{bmk}}$	$\mathbf{Av}(\boldsymbol{\alpha_1^{bmk}})$	$\boldsymbol{\alpha_2^{bmk}}$	$\mathbf{Av}(\boldsymbol{\alpha_2^{bmk}})$
Risk-free rate						
Pos. sig.	1.93%	0.0131	—	—	—	—
Positive	36.52%	0.0035	—	—	—	—
Negative	38.97%	-0.0029	—	—	—	—
Neg. sig.	22.59%	-0.0058	—	—	—	—
Adj. R^2	54.47%	—	—	—	—	—
S&P 500						
Pos. sig.	—	—	2.38%	0.0121	1.27%	0.0146
Positive	—	—	36.39%	0.0036	30.06%	0.0033
Negative	—	—	36.50%	-0.0031	37.98%	-0.0032
Neg. sig.	—	—	24.73%	-0.0072	30.69%	-0.0065
Adj. R^2	—	—	52.87%	—	53.19%	—
Sensoy (2009)						
Pos. sig.	—	—	2.93%	0.0106	1.48%	0.0136
Positive	—	—	36.70%	0.0037	29.10%	0.0032
Negative	—	—	34.66%	-0.0030	35.21%	-0.0033
Neg. sig.	—	—	25.72%	-0.0068	34.21%	-0.0062
Adj. R^2	—	—	53.11%	—	53.50%	—
Modified Sensoy (2009)						
Pos. sig.	—	—	4.20%	0.0093	2.86%	0.0110
Positive	—	—	39.02%	0.0038	34.01%	0.0038
Negative	—	—	33.78%	-0.0028	31.88%	-0.0032
Neg. sig.	—	—	22.94%	-0.0068	31.24%	-0.0063
Adj. R^2	—	—	55.82%	—	55.91%	—
No cash						
Pos. sig.	—	—	4.10%	0.0085	1.77%	0.0120
Positive	—	—	34.41%	0.0035	27.56%	0.0032
Negative	—	—	29.18%	-0.0027	29.76%	-0.0029
Neg. sig.	—	—	32.31%	-0.0060	40.92%	-0.0057
Adj. R^2	—	—	61.60%	—	61.90%	—
Cash adjusted						
Pos. sig.	—	—	6.13%	0.0083	3.86%	0.0108
Positive	—	—	36.14%	0.0038	32.58%	0.0038
Negative	—	—	24.87%	-0.0025	22.30%	-0.0029
Neg. sig.	—	—	32.85%	-0.0056	41.26%	-0.0055
Adj. R^2	—	—	63.15%	—	63.26%	—

Note: This table presents the percentage of funds with significant (5% level, based on two-tailed *t*-statistic) and insignificant estimated alpha under three models (Models (9), (10), and (11)) with five benchmarks (S&P 500, Sensoy (2009) time-invariant benchmarks, no-cash benchmarks, cash-adjusted benchmarks, and cash-augmented benchmarks). The columns labeled α_{rf} and Av(α_{rf}) report the percentage of funds and average of estimated alphas in each group using Model (9). The columns labeled α_1^{bmk} and Av(α_1^{bmk}) report the results using Model (10). The columns labeled $\alpha_2^{bmk}$and Av(α_2^{bmk}) report the results using Model (11).

TABLE 5 Percentage of funds with (in)significant estimated alpha using Fama–French three factors

	Model (9)		**Model (10)**		**Model (11)**	
Benchmark	$\boldsymbol{\alpha_{rf}}$	**Av**$(\boldsymbol{\alpha_{rf}})$	$\boldsymbol{\alpha_1^{bmk}}$	**Av**$(\boldsymbol{\alpha_1^{bmk}})$	$\boldsymbol{\alpha_2^{bmk}}$	**Av**$(\boldsymbol{\alpha_2^{bmk}})$
Risk-free rate						
Pos. sig.	2.07%	0.0132	—	—	—	—
Positive	36.84%	0.0036	—	—	—	—
Negative	38.98%	-0.0029	—	—	—	—
Neg. sig.	22.10%	-0.0058	—	—	—	—
Adj. R^2	53.41%	—	—	—	—	—
S&P 500						
Pos. sig.	—	—	2.34%	0.0121	1.45%	0.0146
Positive	—	—	36.52%	0.0037	30.95%	0.0034
Negative	—	—	38.09%	-0.0031	39.56%	-0.0032
Neg. sig.	—	—	23.05%	-0.0070	28.04%	-0.0063
Adj. R^2	—	—	51.49%	—	51.81%	—
Sensoy (2009)						
Pos. sig.	—	—	2.77%	0.0109	1.54%	0.0140
Positive	—	—	36.91%	0.0038	29.86%	0.0034
Negative	—	—	35.89%	-0.0031	36.59%	-0.0033
Neg. sig.	—	—	24.43%	-0.0068	32.01%	-0.0062
Adj. R^2	—	—	52.27%	—	52.67%	—
Modified Sensoy (2009)						
Pos. sig.	—	—	4.08%	0.0096	2.90%	0.0112
Positive	—	—	39.82%	0.0039	34.82%	0.0039
Negative	—	—	34.25%	-0.0029	32.78%	-0.0031
Neg. sig.	—	—	21.85%	-0.0069	29.51%	-0.0064
Adj. R^2	—	—	54.97%	—	55.07%	—
No cash						
Pos. sig.	—	—	3.81%	0.0091	1.82%	0.0121
Positive	—	—	34.30%	0.0036	27.18%	0.0034
Negative	—	—	29.31%	-0.0027	30.99%	-0.0029
Neg. sig.	—	—	32.58%	-0.0060	40.00%	-0.0057
Adj. R^2	—	—	61.05%	—	61.37%	—
Cash adjusted						
Pos. sig.	—	—	5.78%	0.0086	3.95%	0.0108
Positive	—	—	36.43%	0.0039	33.23%	0.0039
Negative	—	—	24.70%	-0.0025	21.83%	-0.0029
Neg. sig.	—	—	33.10%	-0.0057	40.99%	-0.0055
Adj. R^2	—	—	62.73%	—	62.85%	—

Note: This table presents the percentage of funds with significant (95% confidence, based on two-tailed *t*-

statistics) and insignificant estimated alpha under three models with five benchmarks: (1) Fama–French three-factor model ($r_{it} - rf_t = \alpha_i + \beta_i MKT_t + s_i SMB_t + h_i HML_t + \epsilon_{it}$), (2) three-factor model augmented with ($r_{bmk,k,t} - rf_t - \mathrm{E}(f_t')\lambda_k$), and (3) three-factor model using benchmark-adjusted returns. Variables are defined in Table 3. The columns labeled α_{rf} and $\mathrm{Av}(\alpha_{rf})$ report the percentage of funds and average of estimated alphas in each group using Model (9). The columns labeled α_1^{bmk} and $\mathrm{Av}(\alpha_1^{bmk})$ report the results using Model (10). The columns labeled α_2^{bmk} and $\mathrm{Av}(\alpha_2^{bmk})$ report the results using model (11).

TABLE 6 Percentage of funds with (in)significant estimated alpha using Fama–French five factors

Benchmark	**Model (9)**		**Model (10)**		**Model (11)**	
	$\boldsymbol{\alpha_{rf}}$	$\mathbf{Av}(\boldsymbol{\alpha_{rf}})$	$\boldsymbol{\alpha_1^{bmk}}$	$\mathbf{Av}(\boldsymbol{\alpha_1^{bmk}})$	$\boldsymbol{\alpha_2^{bmk}}$	$\mathbf{Av}(\boldsymbol{\alpha_2^{bmk}})$
Risk-free rate						
Pos. sig.	5.49%	0.0125	—	—	—	—
Positive	36.80%	0.0048	—	—	—	—
Negative	34.67%	-0.0030	—	—	—	—
Neg. sig.	23.03%	-0.0059	—	—	—	—
Adj. R^2	54.53%	—	—	—	—	—
S&P 500						
Pos. sig.	—	—	6.06%	0.0125	3.77%	0.0131
Positive	—	—	36.09%	0.0047	34.84%	0.0049
Negative	—	—	33.48%	-0.0033	34.30%	-0.0033
Neg. sig.	—	—	24.37%	-0.0074	27.09%	-0.0065
Adj. R^2	—	—	52.91%	—	53.24%	—
Sensoy (2009)						
Pos. sig.	—	—	6.65%	0.0118	2.70%	0.0141
Positive	—	—	35.46%	0.0046	32.47%	0.0045
Negative	—	—	32.06%	-0.0032	33.76%	-0.0033
Neg. sig.	—	—	25.82%	-0.0070	31.06%	-0.0064
Adj. R^2	—	—	53.23%	—	53.63%	—
Modified Sensoy (2009)						
Pos. sig.	—	—	8.60%	0.0109	6.03%	0.0117
Positive	—	—	37.32%	0.0047	34.39%	0.0049
Negative	—	—	30.56%	-0.0030	30.95%	-0.0031
Neg. sig.	—	—	23.52%	-0.0070	28.63%	-0.0065
Adj. R^2	—	—	55.78%	—	55.88%	—
No cash						
Pos. sig.	—	—	8.26%	0.0112	4.69%	0.0125
Positive	—	—	33.62%	0.0045	32.71%	0.0046
Negative	—	—	25.07%	-0.0030	24.98%	-0.0030
Neg. sig.	—	—	33.05%	-0.0062	37.63%	-0.0057
Adj. R^2	—	—	62.08%	—	62.44%	—
Cash adjusted						
Pos sig	—	—	10.26%	0.0108	7.90%	0.0114
Positive	—	—	34.01%	0.0047	32.35%	0.0048
Negative	—	—	21.94%	-0.0027	20.42%	-0.0028
Neg sig	—	—	33.78%	-0.0058	39.32%	-0.0055
Adj. R^2	—	—	63.28%	—	63.40%	—

Note: This table presents the percentage of funds with significant (95% confidence, based on two-tailed t-statistics) and insignificant estimated alpha under three models with five benchmarks: (1) Fama–French five-factor model, (2) five-factor model augmented with ($r_{bmk,k,t} - rf_t - \mathrm{E}(f_t')\lambda_k$), and (3) five-factor model using benchmark-adjusted returns. The columns labeled α_{rf} and $\mathrm{Av}(\alpha_{rf})$ report the percentage of funds and average of estimated alphas in each group using Model (9). The columns labeled α_1^{bmk} and $\mathrm{Av}(\alpha_1^{bmk})$ report the results using Model (10). The columns labeled α_2^{bmk} and $\mathrm{Av}(\alpha_2^{bmk})$ report the results using Model (11).

TABLE 7 Out-of-sample forecasting performance using Model (9)

Benchmark	**2003–2006**	**2007–2010**	**2011–2014**
Risk-free rate			
Discrepancy	2.57% (-15.86%)	-0.48% (-27.46%)	5.01% (-17.05%)
First quartile	13.18% (13.99%)	0.58% (3.61%)	11.67% (6.94%)
Second quartile	19.50% (16.99%)	0.72% (4.63%)	15.72% (8.00%)
Third quartile	20.87% (22.22%)	2.22% (14.48%)	10.23% (21.09%)
Fourth quartile	10.61% (29.86%)	1.06% (31.07%)	6.65% (23.99%)
S&P 500			
Discrepancy	7.18% (-20.96%)	-0.49% (-28.20%)	7.72% (-20.17%)
First quartile	15.45% (12.87%)	0.58% (3.31%)	12.90% (5.55%)
Second quartile	20.58% (14.92%)	0.72% (5.01%)	17.23% (7.56%)
Third quartile	18.77% (22.87%)	2.23% (14.24%)	8.57% (21.43%)
Fourth quartile	8.27% (33.83%)	1.07% (31.51%)	5.18% (25.72%)
Sensoy (2009)			
Discrepancy	6.32% (-20.66%)	-0.32% (-27.37%)	9.25% (-18.86%)
First quartile	12.65% (14.78%)	0.58% (3.17%)	13.22% (4.96%)
Second quartile	24.61% (12.52%)	1.01% (5.76%)	17.31% (7.34%)
Third quartile	19.73% (21.52%)	2.10% (14.22%)	9.03% (24.30%)
Fourth quartile	6.33% (35.44%)	0.90% (30.54%)	3.97% (23.82%)
Modified Sensoy (2009)			
Discrepancy	6.08% (-20.87%)	-0.48% (-24.28%)	11.71% (-22.60%)
First quartile	15.22% (14.15%)	0.43% (3.61%)	14.91% (3.52%)
Second quartile	24.04% (12.46%)	0.58% (5.22%)	16.90% (6.11%)
Third quartile	15.86% (21.45%)	2.62% (17.03%)	8.33% (25.33%)
Fourth quartile	9.14% (35.03%)	0.91% (27.90%)	3.20% (26.13%)
No cash			
Discrepancy	10.94% (-29.95%)	-0.05% (-31.62%)	9.46% (-19.05%)
First quartile	14.80% (14.25%)	0.57% (3.27%)	12.72% (5.81%)
Second quartile	26.10% (10.30%)	1.16% (4.91%)	16.69% (8.62%)
Third quartile	17.13% (17.71%)	2.19% (11.66%)	11.30% (20.57%)
Fourth quartile	3.87% (44.20%)	0.62% (34.89%)	3.27% (24.86%)
Cash adjusted			
Discrepancy	17.16% (-34.48%)	0.11% (-29.80%)	13.04% (-23.01%)
First quartile	18.48% (12.23%)	0.58% (3.17%)	14.79% (3.70%)
Second quartile	26.46% (9.19%)	1.31% (4.35%)	18.03% (6.62%)
Third quartile	14.76% (19.54%)	2.14% (14.00%)	8.84% (23.74%)
Fourth quartile	1.32% (46.70%)	0.47% (32.97%)	1.76% (26.71%)

Note: This table presents the percentage of significant positive (negative in parentheses) fund alphas as well as our defined "discrepancy" using Model (9) for three nonoverlapping out-of-sample periods (2003–2006, 2007–2010, 2011–2014). The quartiles are obtained using the *t*-statistic of alpha estimates from Model (11) with the six benchmarks reported in the table, measured over the prior 4 years.

TABLE 8 Cross-section of mutual fund alphas from Models (9) and (12)

	Fama–French–Carhart four-factor model			**CAPM (no cash)**			**CAPM (cash adjusted)**		
		Simula			**Simula**			**Simula**	
Quantile	**Actual**	**ted**	***p*-value**	**Actual**	**ted**	***p*-value**	**Actutal**	**ted**	***p*-value**
Top	4.12%	5.54%	0.78	4.06%	5.22%	0.74	4.46%	5.03%	0.58
2	3.17%	4.13%	0.88	3.97%	4.01%	0.45	4.00%	3.85%	0.36
3	3.14%	3.58%	0.74	2.98%	3.50%	0.83	3.78%	3.34%	0.20
4	3.10%	3.24%	0.56	2.78%	3.21%	0.84	3.30%	3.05%	0.25
5	3.00%	3.00%	0.44	2.71%	2.99%	0.77	2.84%	2.84%	0.44
1%	1.43%	1.31%	0.04	1.23%	1.41%	1.00	1.34%	1.36%	0.59
3%	0.86%	0.85%	0.39	0.72%	0.97%	1.00	0.77%	0.94%	1.00
5%	0.71%	0.67%	0.01	0.56%	0.80%	1.00	0.62%	0.78%	1.00
10%	0.52%	0.44%	0.00	0.37%	0.60%	1.00	0.41%	0.58%	1.00
20%	0.30%	0.24%	0.00	0.13%	0.42%	1.00	0.17%	0.41%	1.00
30%	0.13%	0.13%	0.42	-0.04%	0.32%	1.00	-0.01%	0.32%	1.00
40%	-0.02%	0.06%	1.00	-0.16%	0.25%	1.00	-0.16%	0.25%	1.00
Median	-0.13%			-0.26%			-0.26%		
40%	-0.22%	-0.06%	0.00	-0.34%	0.14%	0.00	-0.34%	0.14%	0.00
30%	-0.30%	-0.13%	0.00	-0.43%	0.07%	0.00	-0.43%	0.08%	0.00
20%	-0.42%	-0.24%	0.00	-0.55%	-0.03%	0.00	-0.55%	-0.02%	0.00
10%	-0.64%	-0.44%	0.00	-0.77%	-0.22%	0.00	-0.77%	-0.21%	0.00
5%	-0.90%	-0.67%	0.00	-1.05%	-0.45%	0.00	-1.05%	-0.43%	0.00
3%	-1.11%	-0.85%	0.00	-1.28%	-0.63%	0.00	-1.28%	-0.60%	0.00
1%	-1.59%	-1.32%	0.00	-1.73%	-1.07%	0.00	-1.84%	-1.02%	0.00
5	-3.42%	-2.96%	0.11	-3.93%	-2.48%	0.00	-4.05%	-2.39%	0.00
4	-3.54%	-3.20%	0.20	-3.95%	-2.67%	0.01	-4.08%	-2.56%	0.00
3	-3.68%	-3.52%	0.34	-4.11%	-2.93%	0.03	-4.28%	-2.82%	0.01
2	-3.95%	-4.05%	0.47	-4.37%	-3.39%	0.10	-5.34%	-3.29%	0.02
Bottom	-4.39%	-5.24%	0.65	-5.52%	-4.47%	0.19	-5.35%	-4.27%	0.16
Adj. R^2	54.47%			60.24%			61.87%		

Note: This table reports risk-adjusted monthly alphas estimated using the Fama–French–Carhart four-factor model and capital asset pricing model (CAPM) based on our time-varying indices-based benchmarks for both actual and simulated mutual funds, ranked from highest (top) to lowest (bottom). The first two columns report the selected quantiles and the cumulative distribution function (CDF) of the actual estimated alphas at selected quantiles when they are ranked from the highest to lowest, and the following two columns report the CDF of the simulated "luck" distribution as well as the *p*-values that correspond to the selected quantiles of the distribution of the simulated alphas based on Model (9). Analogically, the remaining six columns report the results based on CAPM with our proposed no-cash benchmarks and cash-adjusted benchmarks, respectively.

TABLE 9 Cross-section of *t*-statistics of mutual fund alphas

	Fama–French–Carhart four-factor model			CAPM (no cash)			CAPM (cash adjusted)		
		Simulate			Simul		Actua	Simu	
Quantile	Actual	d	*p*-value	Actutal	ated	*p*-value	l	lated	*p*-value
Top	6.69	5.04	0.06	6.00	19.39	1.00	9.71	31.56	1.00
2	5.50	4.39	0.02	5.79	15.86	1.00	9.70	24.24	1.00
3	5.31	4.08	0.01	5.73	14.26	1.00	7.54	21.46	1.00
4	5.29	3.90	0.00	5.27	13.33	1.00	7.18	19.59	1.00
5	4.52	3.77	0.01	4.91	12.68	1.00	6.70	18.16	1.00
1%	2.31	2.58	1.00	2.32	7.22	1.00	2.52	9.67	1.00
3%	1.79	2.04	1.00	1.53	5.48	1.00	1.71	6.78	1.00
5%	1.53	1.77	1.00	1.23	4.74	1.00	1.38	5.59	1.00
10%	1.16	1.37	1.00	0.80	3.70	1.00	0.94	4.14	1.00
20%	0.72	0.89	1.00	0.33	2.59	1.00	0.43	2.78	1.00
30%	0.33	0.55	1.00	-0.11	1.93	1.00	-0.04	2.03	1.00
40%	-0.07	0.27	1.00	-0.62	1.45	1.00	-0.63	1.51	1.00
Median	-0.51			-1.22			-1.29		
40%	-0.94	-0.27	0.00	-1.98	0.68	0.00	-2.22	0.72	0.00
30%	-1.51	-0.55	0.00	-2.98	0.31	0.00	-3.33	0.34	0.00
20%	-2.14	-0.89	0.00	-4.27	-0.10	0.00	-4.71	-0.07	0.00
10%	-2.99	-1.36	0.00	-6.12	-0.64	0.00	-6.59	-0.61	0.00
5%	-3.71	-1.76	0.00	-7.71	-1.07	0.00	-8.10	-1.05	0.00
3%	-4.23	-2.03	0.00	-8.64	-1.35	0.00	-9.14	-1.33	0.00
1%	-5.13	-2.57	0.00	-11.03	-1.89	0.00	-11.38	-1.87	0.00
5	-7.29	-3.76	0.00	-16.13	-2.98	0.00	-16.17	-2.94	0.00
4	-7.34	-3.91	0.00	-16.51	-3.09	0.00	-17.80	-3.06	0.00
3	-7.40	-4.10	0.00	-17.76	-3.23	0.00	-17.86	-3.20	0.00
2	-7.99	-4.39	0.00	-21.05	-3.46	0.00	-18.57	-3.41	0.00
Bottom	-9.21	-5.08	0.01	-23.67	-3.99	0.00	-20.05	-3.92	0.00

Note: This table reports *t*-statistics of risk-adjusted monthly alphas estimated using the Fama–French–Carhart four-factor model and capital asset pricing model (CAPM) based on our time-varying indices-based benchmarks for both actual and simulated mutual funds, ranked from highest (top) to lowest (bottom). The first column reports the cumulative distribution function (CDF) of the actual estimated *t*-statistics at selected quantiles, and the following two columns report the CDF of the simulated "luck" distribution as well as the *p*-values that correspond to the selected quantiles of the distribution of the simulated *t*-statistics based on Model (9). Analogically, the remaining six columns report the results based on CAPM with our proposed no-cash benchmarks and cash-adjusted benchmarks, respectively. The *t*-statistics of alpha are based on heteroskedasticity- and autocorrelation-consistent standard errors.