Monthly Measurement of Daily Timers[0]

William N. Goetzmann
Yale School of Management

Jonathan Ingersoll Jr.
Yale School of Management

Zoran Ivković
Yale School of Management

April 25, 1999

[0] We thank the anonymous referee for constructive suggestions. We also thank Don Chance, Steve Shellans, and Ed Owens for useful discussions on this topic. All errors are the sole responsibility of the authors.

Monthly Measurement of Daily Timers

Abstract

We examine the power of the Henriksson-Merton test through simulations of a market and a timer who can move in and out of the market on a daily basis. Our simulations show that the Henriksson-Merton measure is weak and biased downward when applied to the monthly returns of a timer who makes choices daily. We propose a simple solution that alleviates the problem without collecting daily timer returns. The solution uses daily returns to an index correlated to the timer's risky asset: values of a daily put on the index are cumulated over each month to form a regressor that captures timing skill. Our simulations indicate that the adjusted HM measure applied to monthly returns is much more powerful and reduces the bias in the estimated put value. Next, we study four tests of timing skill: the classic HM test, our adjusted test, the HM-FF3 test (a modification of the classic HM test which utilizes the Fama-French 3-factor model instead of CAPM), and our adjusted-FF3 test (a modification of our adjusted test which utilizes the Fama-French 3-factor model). Very few funds from our sample of 558 mutual funds exhibited statistically significant positive timing skill under either measure. More encompassing, our empirical analyses indicate that the adjusted-FF3 test is the least biased measure of timing skill among the four. The adjusted-FF3 test improves upon the classic HM test in two important ways. First, the adjusted timing instrument provides for a sharper inference regarding timing skill. Second, the choice of the Fama-French 3-factor model helps mitigate biases associated with the choice of investment style. Finally, we document the impact of survivorship by comparing the results obtained for surviving funds to those obtained for defunct funds.

I. Introduction

Henriksson and Merton (Merton, 1981; Henriksson and Merton, 1984) [HM] develop a logically appealing measure of market-timing skill. Their analysis is based upon the simple intuition that a market timer effectively provides a put to the client. When the market is up, the perfect timer is fully invested in the risky asset. When the market is down, the perfect timer will be holding the riskless asset. HM show how a simple parametric test — a regression of portfolio returns on two variables — can be used to estimate a manager's timing skill. In this paper, we focus on the problem of using the test on monthly data when managers make daily timing decisions.

Several researchers have studied the HM timing measure.[1] Glosten and Jagannathan (1994) show it to be a special case of a more general contingent claims approach to performance evaluation. They propose an improved version of the HM test that allows for managed portfolios to represent bundles of multiple options with different strikes. Implicit in the empirical application of their framework, however, is the presumption that the options share a common maturity. This is the heart of the problem we address. Not only is the timer effectively holding or mimicking a bundle of options with varying strikes, these options are effectively being rolled over at a frequency potentially unequal to the interval of return estimation. Only one paper to our knowledge points out the magnitude of this monthly data problem. Chance and Hemler (1998) strongly reject the null of no timing ability for a manager using daily data but find that all evidence of timing ability disappears when monthly data are used. In this paper, we show that the use of monthly data essentially implies that most standard timing tests are misspecified; it should thus come as little surprise that few researchers to date have found evidence of timing ability by professional managers.

In general, evidence on the ability of investment managers to time the market is mixed. Several studies of mutual fund timing skill, e.g., Treynor and Mazuy (1966), Henriksson (1984), Chang and Lewellen (1984), and Grinblatt and Titman (1989a), generally found little evidence of timing skill. On the other hand, Ferson and Schadt (1996) found some evidence of manager timing skill when macroeconomic conditions are accounted for; Graham and Harvey (1996) found evidence of timing skill using certain benchmarks. Wagner, Shellans and Paul (1992), Brocanto and Chandy (1994), and Chance and Hemler (1998) all found some positive timing evidence as well. Brown, Goetzmann, and Kumar (1997) found evidence that the Dow Theory worked as a timing strategy. While our study focuses specifically on a correction for the HM parametric test of timing skill, it may generally be the case that at least some of the ambiguity in the existing results is due to the fact that most existing studies relied upon monthly returns. It also has direct implications for the tests proposed by Glosten and Jagannathan (1994); allowing for more frequent timing activity may improve the power of their tests.

This paper is organized as follows. The next section discusses the Henriksson-Merton parametric test of timing skill, defines our adjusted measures of timing skill, and sets the stage for simulation and empirical analyses reported in subsequent sections. Section III describes our simulations and the simulation results. Section IV describes empirical results. Section V concludes.

[1] For an excellent review see Grinblatt and Titman (1995).

II. Henriksson-Merton Tests of Timing Skill

II.1. Development

In their 1981 paper, Henriksson and Merton develop two tests of timing skill. One is a non-parametric test that relies upon knowing the timer's forecast of the market.[2] The other is a parametric test that relies solely on the returns generated by the timer. For cases in which the timer's forecast is known, the non-parametric test is a direct test of the timer's forecasting skill. In most circumstances, however, the timer's forecast is unknown. Few investment managers report their forecasts as well as their performance. Tests of timing ability of mutual fund managers, for example, typically rely upon monthly fund returns.

The Henriksson-Merton parametric test is a linear regression of the timer's portfolio excess returns $Z_{p,t} \equiv R_{p,t} - R_{f,t}$ ($R_{.,t}$ denotes net returns in period t) on the constant and two variables:

$$Z_{p,t} = \alpha + \beta Z_{m,t} + \gamma \max\{-Z_{m,t}, 0\} + \epsilon_t. \tag{1}$$

The first variable, $Z_{m,t} \equiv R_{m,t} - R_{f,t}$, is the excess return on the risky asset (i.e., the market) and the second variable captures the value of the implicit protective put. It takes on the value 0 when the excess return of the market is positive and it exactly offsets losses when the market drops by taking the value $-Z_{m,t}$. A perfect pure market timer should have a market coefficient β of one and a timing coefficient γ of one. This corresponds to a long position in the asset and a long position in a put with a maturity of one period struck at-the-money at the beginning period asset price. Notice that this formulation implicitly assumes that the market timer will either be in the market over the entire period or out of the market over the entire period. That is, in terms of its systematic risk, the pure timer's portfolio beta toggles between values of one and zero.

The real-world HM-style timer's strategy would likely be less aggressive and would instead be limited to toggling between a high beta, β_h, and a low beta, β_l, in anticipation of a bull market (i.e., $Z_{m,t} > 0$) and a bear market (i.e., $Z_{m,t} \leq 0$), respectively. Fortunately, the HM model readily captures such a perfect timer: it is recognized that the timing coefficient for a perfect timer in the above regression will be better represented by the difference between the two betas, i.e., $\gamma = \beta_h - \beta_l$.

[2] Pesaran and Timmermann (1994) show that the non-parametric test is equivalent to a Fisher's exact test about a 2×2 matrix.

Of course, a real-world HM-style timer would not generate perfect forecasts. Merton (1981) defined the conditional probability $p_1(t)$ of a correct forecast at time t given that $Z_{m,t+1} \leq 0$ and the conditional probability $p_2(t)$ of a correct forecast at time t given that $Z_{m,t+1} > 0$. The timing coefficient for a timer who generates forecasts with such accuracy is $\gamma = (p_1 + p_2 - 1)(\beta_h - \beta_l)$. Intuitively, this value of gamma indicates that the timer's forecasts have positive value if both $p_1 + p_2 > 1$, i.e., if the timer generates "good" forecasts, and $\beta_h - \beta_l > 0$, i.e., if the timer reacts to the forecasts appropriately.

On the other hand, a HM timer is completely oblivious to the magnitude of the anticipated excess return. For example, even the perfect HM timer would be in the market with the same beta of one (or that "high" beta β_h which is consistent with the manager's investment strategy) both if the anticipated excess return in the next period, known to the perfect HM timer with certainty, is barely positive and if it is very large. In other words, the HM model does not allow the HM timer's systematic portfolio risk to vary with the timing signal in any but the most restrictive way. This criticism was addressed in the literature; it was one of the motivating factors that prompted Admati, Bhattacharya, Pfleiderer, and Ross (1986) to consider a framework in which the portfolio beta is a function of the timing signal. Specifically, under the assumptions of exponential utility and multivariate normality of asset returns, Admati, Bhattacharya, Pfleiderer, and Ross (1986) showed that the timer's portfolio beta is related to the timing signal in a linear fashion. Moreover, they show that, under the same assumptions, the contribution of timing to the overall performance can be detected via Treynor-Mazuy quadratic regression (Treynor and Mazuy, 1966) as the product of the regression coefficient of the quadratic term and the variance of market index used to construct the regressors. More recently, Ferson and Schadt (1996) studied the conditional version of the HM model. Their model allows the portfolio beta to vary with several macroeconomic variables that have previously been shown to have some predictive power to forecast future market returns.

At the intuitive level, both the HM specification and the Treynor-Mazuy specification rely on the premise that a successful timer will adjust the portfolio's systematic risk by increasing/decreasing it in anticipation of a bull/bear market *and* that there is no co-skewness between the assets held in the portfolio and the benchmark. Clearly, co-skewness will contribute toward a possibly incorrect finding that the manager possesses timing "ability." Unfortunately, it has long been known (Kraus and Litzenberger, 1976) that many stocks are co-skewed with market returns. Furthermore, Jagannathan and Korajczyk (1986) argue that portfolio co-skewness can be induced by pursuing dynamic portfolio strategies, e.g., by buying call options on the market or by buying small, highly levered stocks. According to

Jagannathan and Korajczyk (1986), small stocks exhibit option-like characteristics that can induce spurious positive timing ability. Interestingly, Low (1999) showed that small stocks exhibit negative timing characteristics when both beta and covariation with a bullish market are controlled for. In summary, it appears that HM-style tests (as well as Treynor-Mazuy-style tests) can be gamed (purposely or not).

Measuring performance in the presence of timing is inherently difficult. It has long been known that many mutual funds exhibit returns that are nonlinearly related to index returns (see, e.g., Lehmann and Modest, 1987). Thus, there are indications that many managers exert at least some attempts of timing, i.e., of active management beyond stock picking (selectivity). While classic performance measures, e.g., Jensen's alpha (Jensen, 1968, 1969) have been shown to be biased in the presence of timing activity (see, e.g., Grinblatt and Titman, 1989b) and are thus inadequate measures of the overall performance, separating selectivity and timing performance as was proposed in the HM model (and, similarly, in the TM model) is not entirely immune to bias. Recently, Kothari and Warner (1997) carried out a detailed study of standard mutual fund performance measures, including the HM measure. They created simulated portfolios by randomly picking stocks (sometimes controlling for size or book-to-market ratio) and periodically changing the portfolio composition so as to mimick the turnover of a typical mutual fund. While such portfolios are clearly not derived from skill, neither in selection nor in timing, standard performance measures (including the HM measure) nevertheless detected abnormal performance. Kothari and Warner concluded that "... the performance measures [studied in their paper] are badly misspecified" (Kothari and Warner, 1997, p. 2). Furthermore, selectivity and timing measures seem to be confounded. Negative correlation between the two in the context of the HM model (and beyond) has been reported by Kon (1983), Henriksson (1984), and Jagannathan and Korajczyk (1986). Interestingly, Pfleiderer and Bhattacharya (1983) noticed that negative correlation between measures of timing and selectivity could be induced by intraperiod trading.

An alternative approach to return-based performance evaluation is to design methods that estimate measures of overall performance, i.e., measures that simultaneously capture selectivity and timing. Prominent examples include Grinblatt and Titman's PPW (Positive Period Weighting) measures (Grinblatt and Titman, 1989b; Cumby and Glen, 1990; Grinblatt and Titman, 1994), Glosten and Jagannathan's contingent claims approach (Glosten and Jagannathan, 1994), and a variety of techniques proposed by Chen and Knez (1996).[3]

[3]Yet another approach relies on detailed information on portfolio weights. See, e.g., Grinblatt and Titman (1989a, 1993) for details.

Despite criticisms and limitations, the HM measure and its generalizations remain among the most frequently used methods of performance evaluation. Glosten and Jagannathan (1994) recognize the HM measure as an important special case of their more general contingent claims approach to performance evaluation. Ferson and Schadt (1996) extend the HM framework to a conditional setting in which the portfolio beta is a linear function of the unexpected changes in a set of pre-specified macro-economic variables. Both of these researches represent significant advances in lessening the impact of restrictive behavioral assumptions imposed by the original HM method while retaining its intuitive appeal, relative ease of implementation, and minimal data requirements.

II.2. Contribution

Our research looks at yet another (possibly severe) behavioral restriction of the original HM model — the assumption that trading frequency and return measurement frequency are identical. In fact, if a timer could trade within the period, then Equation 1 is misspecified because the variable capturing the value of the implicit put does not account properly for the value of intermediate investment decisions. Intuitively, in an up month for the market even a moderately successful daily timer should generate a return that exceeds the market return. However, that success will not necessarily be credited to timing skill; the value of the timing instrument $\max\{0, -Z_{m,t}\}$ will be zero and it will thus be impossible to distinguish how much of the timer's performance is due to timing skill.

This problem is exacerbated as the difference between the decision horizon and the evaluation horizon grows. Many investment managers report only quarterly performance. As the horizon grows, the frequency of negative period returns for the risky asset decreases, and so does the power of the HM parametric test, which relies upon the covariance of the manager returns with the put value conditional upon the risky asset underperforming. Consequently, as our simulations will show, the HM test for daily timers using monthly data is extraordinarily weak.

The best solution to the problem is to collect data that corresponds to the frequency with which the timer makes decisions. This is typically not possible. While Busse (1997) collects daily mutual fund data for investigating whether mutual fund managers in general time the variance of the market and Chance and Hemler (1998) have daily data for a limited number of market-timers, almost no money manager reports daily results in a form generally accessible to researchers and analysts. An alternative to collecting daily data is to collect daily data on the risky asset alone. Daily S&P 500 returns, for example, can be used to construct an instrument that is correlated to the daily put values. More precisely, we cumulate the value of the daily puts over the month to estimate the monthly value of a daily timer's skill. We define the **adjusted** test as follows:

$$Z_{p,t} = \alpha + \beta Z_{m,t} + \gamma P_{m,t} + \epsilon_t, \quad P_{m,t} = \left[\left(\prod_{\tau \in month(t)}^{t} \max\{1 + R_{m,\tau}, 1 + R_{f,\tau}\}\right) - 1\right] - R_{m,t}, \; (2)$$

where $P_{m,t}$ is the value added by perfect daily timing per dollar of fund assets. Even when daily returns on the risky asset timed by the timer are not available, as long as the asset returns used by the econometrician to construct the instrument $P_{m,t}$ are highly correlated with them, this specification provides an improvement over the standard HM specification from Equation 1.

In sum, we substitute the value of a monthly put on the market with a rolling account through the month of the gains to having a sequence of daily market puts. Such a sequence of option contracts is often called *tandem options* (Blazenko, Boyle, and Newport, 1990). Blazenko *et al.* (1990) define tandem options as a sequence of options that are "... regularly brought back 'to the money'" (Blazenko *et al.*, 1990, p. 40), i.e., the exercise price of the option is periodically reset to the current asset price. In our context, the exercise price is reset daily to the value that equals the product of the current value of the risky asset and the gross daily return on the riskless asset which prevails on that day.

The cumulation of daily puts necessitates a behavioral assumption about the strategy pursued by the perfect daily timer. Every day, the perfect daily timer will take the proceeds from the payoff from the daily put option that expires on that day (if it expired in-the-money) and invest it in the same way as the remainder of the portfolio. That is, if the timer forecasts a positive excess return, the timer takes a 100% position (investing both "old" funds and the newly acquired payoff from the daily put option) in the risky asset. Conversely, if the timer forecasts a negative excess return, the timer takes a 100% position in the riskless asset. This behavioral assumption fully conforms with the notion that the least a perfect daily timer could do with the proceeds from the daily put is to invest it in the riskless asset and thus earn zero excess return. However, having perfect foresight, the perfect timer can do even better — to seek positive excess return even for the investment of the proceeds from the daily put whenever the forecast indicates a positive excess return and to resort to the riskless investment otherwise. Put differently, any value generated from daily puts will generate at least the riskless rate of return from the day proceeds are collected onward, and will do even better each time a positive excess return is forecast for the day.

At the conclusion of this section we turn our attention to the issue of the underlying asset pricing model. Both the classic HM test from Equation 1 and the above adjusted test from Equation 2 are based on the classic Sharpe-Lintner-Mossin CAPM (Sharpe, 1964; Lintner, 1965; Mossin, 1966). CAPM itself and its use in performance measurement have been subjected to strong objections from the theoretical standpoint (see, e.g., Roll, 1978, 1979; Mayers and Rice, 1979; Admati and Ross, 1985; Dybvig and Ross, 1985). Empirical studies have uncovered risk factors other than the market that are relevant in explaining cross-sectional variation in average asset returns and are thus questioning the validity of the CAPM. Among those, size and book-to-market ratio have been extensively studied (Banz, 1981; Rosenberg, Reid, and Lanstein, 1985; Fama and French, 1992, 1993, 1996), and a multi-factor asset pricing model which, in addition to the market, includes risk factors that account for size and for the book-to-market ratio has been proposed by Fama and French (1992, 1993, 1996) and widely accepted by academics and practitioners alike. The three-factor model of Fama and French (1992, 1993, 1996), or indeed any plausible multi-factor asset pricing model can be readily utilized instead of the CAPM; Merton's (1981) analysis is robust to the choice of the underlying asset pricing model. However, the HM test specification from Equation 1 (originally developed in Henriksson and Merton, 1981) would have to change if CAPM were replaced with another asset pricing model. Following the approach from Kothari and Warner (1997), we also carry out the HM test based on the Fama-French 3-factor model (Fama and French, 1992, 1993, 1996):

$$Z_{p,t} = \alpha + \beta_1 Z_{m,t} + \gamma \max\{-Z_{m,t}, 0\} + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \epsilon_t, \tag{3}$$

where SMB_t and HML_t are the returns in month t to the Fama-French size factor and the book-to-market factor zero-cost portfolios, respectively (Fama and French, 1992, 1993, 1996). We will henceforth refer to the regressions based on the specification from Equation 3 as the **HM-FF3** test.

Finally, we define the **adjusted-FF3** test as follows:

$$Z_{p,t} = \alpha + \beta_1 Z_{m,t} + \gamma P_{m,t} + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \epsilon_t, \tag{4}$$

where $P_{m,t}$ is the same instrument as before:

$$P_{m,t} = \left[\left(\prod_{\tau \in month(t)}^{t} \max\{1 + R_{m,\tau}, 1 + R_{f,\tau}\} \right) - 1 \right] - R_{m,t}$$

The motivation for the adjusted-FF3 test specification from Equation 4 parallels the earlier discussion that lead to the development of the adjusted test specification from Equation

2. Simply put, the adjusted-FF3 specification combines the measurement of the monthly value of a daily timer's skill via $P_{m,t}$ with the Fama-French 3-factor asset pricing model. Results obtained by Kothari and Warner (1997) show that biases in performance measurement via the Henriksson-Merton model are smaller if the Fama-French 3-factor asset pricing model is used instead of the CAPM. That is, the specification from Equation 3 is better than the specification from Equation 1. Following that logic, the specification from Equation 4 should be the best among the four specifications, i.e., it should have the most power to detect timing skill and the smallest bias. We will revisit this issue in Section IV.

III. Simulations

We conduct simulations to examine the performance of the HM-style parametric test by employing the classic HM test on daily returns and on monthly returns (both according to specification from Equation 1), and the adjusted test (according to specification from Equation 2).

For each of these tests we report the mean values for coefficients α, β, γ and the frequency with which the null hypothesis of no timing skill is rejected for differing levels of timing skill.

We also run another popular test of timing skill, the Treynor-Mazuy (1996) test and compare its power, both for daily and monthly data, to that of its HM counterparts (Equation 1) and to the power of the adjusted test (Equation 2). The Treynor-Mazuy [TM] test is specified as follows:

$$Z_{p,t} = \alpha + \beta Z_{m,t} + \gamma Z_{m,t}^2 + \epsilon_t.$$

III.1. Simulating the Market

For each simulation, we generate ten years (2,520 days) of daily excess returns to the risky asset (which plays the role of the market) as i.i.d. random variables with an annualized mean of 10% and an annualized standard deviation of 16%. These parameters are characteristic of broadly diversified stock market indexes (consisting of mostly large stocks) in the U.S. capital markets. The generated excess returns are exponentiated (after appropriate correction of the mean) to create lognormal excess returns.

III.2. Simulating the Timer

In order to capture the restrictions associated with mutual fund investing, the simulated HM-style timer is not allowed to take short positions. On any day, the timer can be either fully invested in the risky asset or fully invested in the riskless asset. While the former may yield either positive or negative excess return for the day, the latter always (by definition) yields zero excess return. The simulated HM-style timer forecasts returns to the risky asset on the next day. If the timer's forecast indicates a positive excess return, the timer takes a 100% position in the risky asset; if the forecast indicates a negative excess return, the timer takes a 100% position in the riskless asset.

We define perfect timing skill as the ability to forecast the sign of the excess return to the risky asset on the next day with no error. Thus, the perfect HM-style timer would take a position in the riskless asset if and only if the excess return to the risky asset will be negative. Clearly, generating such forecasts without errors is a tall order; a real-world manager is far more likely to be only moderately successful. Therefore, there is an obvious need to define imperfect timing skill in this context.

We define *skill* as the probability of correctly forecasting at time t the sign of excess return $Z_{m,t+1}$. This definition of skill is in fact a special case of corresponding definitions furnished by Merton (1981). Merton's model allowed for a differentiation between the conditional probability $p_1(t)$ of a correct forecast at time t given that $Z_{m,t+1} \leq 0$ and the conditional probability $p_2(t)$ of a correct forecast at time t given that $Z_{m,t+1} > 0$. In our simulations, we set $skill = p_1 = p_2$. Merton's classic result, which indicates that the manager's forecast has positive value if and only if $p_1 + p_2 > 1$ (Merton, 1981), translates into $skill > 0.5$ in the present framework of daily measurement of timing ability of HM-style daily timers.

Put differently, timing skill can be viewed as a parameter, ranging from *skill*=0 to *skill*=1, that defines the fraction of correct forecasts. The skill level *skill*=1 indicates that the manager correctly forecasts the sign of excess return 100% of the time, i.e., that the manager has perfect timing ability. Conversely, the skill level *skill*=0 indicates that the manager's forecasts are always incorrect, i.e., that the manager has perfect *perverse* timing ability. Particularly interesting is the skill level *skill*=0.5 — each daily forecast is as likely to be correct as it is likely to be incorrect. Finally, note that, consistently with the Henriksson-Merton framework, probabilities of correct forecast do not depend on the magnitude of the excess return, neither do they vary across time (thus eliminating the notion of learning, i.e., improving skills with experience).

We consider various levels of timing skill — from *skill*=0 to *skill*=1 with a step size of 0.1. For each skill level we run 1,000 simulations. In each simulation run, excess returns on the market are simulated in the manner described in Section III.1. Timer's returns are simulated on the basis of 2,520 flips of a biased coin (once for each simulated day). Each coin flip is implemented as the comparison between a pseudo-random draw from the standard uniform distribution and *skill*: if the pseudo-random draw generated for day t exceeds the threshold *skill*, then the forecast is labeled as incorrect and the timer takes the wrong position (i.e., full investment in the riskless asset if $Z_{m,t+1} > 0$ and full investment in the risky asset if $Z_{m,t+1} \leq 0$); if, on the other hand, the pseudo-random draw generated for day t does not exceed the threshold *skill*, then the forecast is labeled as correct and the timer takes the appropriate position (i.e., full investment in the risky asset if $Z_{m,t+1} > 0$ and full investment in the riskless asset if $Z_{m,t+1} \leq 0$).[4]

III.3. Simulation Results

III.3.A. Mean Values

Table 1 reports for each skill level the mean estimated values for the coefficients α, β, and γ over 1,000 simulations of the market for both daily and monthly HM specifications and for the adjusted specification, respectively.

Table 1 about here

The bottom panel of Table 1 reports the mean coefficients for the timing coefficient γ. Notice that the test results based on monthly sampled data exhibit a strong downward bias in the timing coefficient. A perfect timer (*skill*=1) should have a γ coefficient of 1 for both daily and monthly HM specification, but the mean value for monthly sampled data is only about 0.18. This downward bias is consistent with the fact that the effective put is measured with error. The adjusted test, on the other hand, is unbiased for the perfect timer. Since we would not expect to find timers with anything other than modest ability (due to the efficient

[4] We also implemented an alternative simulation in which the timer's forecast of the sign of excess return $Z_{m,t+1}$ was defined as the sign of the mixture of $Z_{m,t+1}$ and a random draw from the distribution of excess returns $Z_{m,t'}$, $t' \in \{1, 2, \ldots, 2520\}$, i.e., $sign(skill \times Z_{m,t+1} + (1 - skill) \times Z_{m,t'})$, $t' \in \{1, 2, \ldots, 2520\}$. One potential concern with this approach is that the variance of the imperfect forecast of the excess return is lower than the variance of either the perfect timer or the no foresight timer because the mixing procedure effectively creates a portfolio. Nevertheless, the simulation results (not reported here) led to conclusions that were identical to those reported in Section III.3.

market theory) this is troublesome evidence – even with perfect data (i.e., a fictitious perfect timer) the estimated value of the put provided by the (perfect) timer under the monthly HM specification is below its true value of one.

The magnitude of the coefficient γ is useful for more than a hypothesis test about timing ability. Since a value of one corresponds to the timer effectively providing a whole put on the equity position, any value less implies that the timer is only providing a partial put. Thus, above and beyond the simple question of whether there is timing skill, the downward bias in the coefficient leads to an incorrect inference about the value added by the manager's timing skill.

The top panel in Table 1 reports the mean α values. Notice that the HM specification using the monthly data appears to attribute the skill (*skill* > 0.5) to positive alphas. This is consistent with the evidence, reported in Kon (1983), Henriksson (1984), and Jagannathan and Korajczyk (1986), that the timing and selection measures are negatively correlated. Were we to use this test on a mutual fund manager, for example, we might infer that the manager had no timing ability, yet had displayed superior selection ability. The monthly adjusted alphas are close to zero, which suggests that the adjusted timing test is not biased towards finding either positive or negative selection ability.

The panel with β values indicates that market risk exposure for the timer with no skill is effectively the average of the exposures over the interval. That is, under all three specifications the mean value of the β coefficient for *skill*=0.5 is about 0.5. When the HM specification is applied to daily returns or adjusted specification is used, the regression appears to successfully distinguish market exposure and timing activity as the timing ability increases. This is only marginally true for the HM specification applied to monthly returns — even perfect daily market timers have a beta of only about 0.63.

All three panels together paint a very distorted image of a perfect daily market timer under the classic monthly HM specification. Instead of receiving due recognition of their timing ability, perfect daily timers are credited with an enormous alpha of over eight percent per month, a beta of 0.6352, and a modest gamma of 0.1808. A similar pattern, albeit on a lesser scale, can be detected for able (but less than perfect) daily timers, i.e., those with a skill level of *skill*=0.6, ..., 0.9. By contrast, the adjusted test is fair in that it appears to give credit where it is due — the range of alphas is far more modest than that of the monthly specification (typically up to eight basis points per month, except for perverse timers with *skill* < 0.4) and gammas appear to better reflect the value of the implicit put provided by the daily timer.

III.3.B. Power

Table 2 focuses on the power of tests about the timing coefficient γ under the three HM-style specifications. The table reports the quantile of the critical t-value of 1.96 on the timing coefficient in the distribution of t-values generated by 1,000 simulations for each skill level and for each of the specifications. The column under *skill*=0.5 reports results under the null hypothesis of no timing skill; the value of 0.879 for the daily HM specification indicates that the null hypothesis would have been falsely rejected about 12% of the time using the traditional 95% confidence level. Interestingly, the columns of Table 2 associated with skill level of *skill*=0.5 also suggests that the power of the daily HM test to identify managers who, within the framework of our simulation model, are completely devoid of timing skill is less than that of both the monthly HM test and the adjusted test. Table 2 also suggests that the power of different specifications of the test to detect timing skill varies dramatically with skill level. For example, the value of the entry in the second row and the last column of Table 2 is 0.570, which suggests that the null hypothesis of no timing skill is rejected only about 43% of the time for a daily timer with perfect foresight (i.e., skill level of *skill*=1) when the standard monthly HM specification is used. The adjusted test displays a dramatic increase in power as the skill level rises from *skill*=0.5 toward *skill*=1; it virtually never fails to reject the null when the skill level is *skill*=0.8 or above. Even when the skill level is *skill*=0.6, the adjusted timing test has some power to reject.

Table 2 about here

Finally, we ran the same simulation using the TM model on both daily and monthly data. The last two rows of Table 2 show the the quantile of the critical t-value of 1.96 on the timing coefficient for both TM-style specifications. A comparison of the first two rows (representing the HM daily and monthly tests) to the last two rows (representing the TM daily monthly tests) of Table 2 reveals a striking finding that the powers of the respective daily and monthly tests are very similar for all skill levels. Thus, the same biases detected for the HM monthly test simulations exist for the TM monthly test simulations. This suggests that a more significant improvement in the power of performance measurement can be accomplished by adjusting for the cumulated value of timing within a month than by choosing another model of timing skill and employing it on monthly data.

IV. Empirical Results

In this section, we apply the four tests from Section III to a set of monthly open-end mutual fund returns. We summarize the results of our analyses for each specification by the funds' Morningstar Category classification. More encompassing, we perform a comparison of the four test specifications on a set of passive stock indexes (which should not exhibit any selection or timing ability) and conclude that the adjusted-FF3 specification (Equation 4) appears to be superior to the other three.

IV.1. Preliminaries

Intuitive appeal of the adjusted timing test from Equations 2 and 4, relative ease of implementation, and the simulation results reported in Section III call for an empirical investigation. Several questions are of interest.

First, do our adjusted timing measures find evidence of timing skill? To what extent, if any, will the assessment of timing skill provided by the adjusted timing tests differ from the existing results based on HM tests?

Second, in light of the simulation evidence that the adjusted tests have greater power, do they provide sharper inference? How will the cross-sectional distributions of the timing coefficients under the HM specifications and our adjusted specifications look like relative to one another?

Third, what is the relation between the performance measurements of a fund and the Morningstar Category the fund belongs to? For example, will some categories feature many funds with selection ability and/or timing ability while other categories have virtually none? Alternatively, will the finding of positive selection ability be routinely accompanied by the finding of negative timing ability and *vice versa*? Are any of these findings really due to managerial skill, apparently shared among most funds from the particularly (un)successful category, or are they an artifact of the misspecification of the risk-return model?

Fourth, many mutual funds neither profess to be timers nor attempt to engage in timing. The HM measure and the adjusted measure will nonetheless produce the timing coefficient which might spuriously indicate positive or negative timing skill. In addition to the point that the measure produces spurious timing (in)ability, another question begs to be asked: do the manager who actually time the market have any true timing skill?

IV.2. Data Description

Monthly total returns for the open-end mutual funds were obtained from the April 1998 Morningstar Principia CD-ROM. Morningstar does not adjust the total returns for sales charges (e.g., front-end charges, deferred fees, redemption fees), but it does account for management fees, administrative fees, 12b-1 fees, and other costs that are automatically taken out of fund assets.

In order to be included into our sample, each fund that was reported on the April 1998 Morningstar disc had to meet all of the following criteria:

1. The fund has to hold at least some stock,
2. Foreign stocks can account for at most 20% of equity holdings held by the fund,
3. The inception date of the fund date should be December 1987 or earlier, and
4. The fund should belong to one of the following ten Morningstar Categories: Large Value, Large Blend, Large Growth, Mid-Cap Value, Mid-Cap Blend, Mid-Cap Growth, Small Value, Small Blend, Small Growth, and Domestic Hybrid.[5]

A total of 558 funds met the above criteria. The criteria were set up with the intent of obtaining a substantial cross-section of funds (hence a relatively recent inception date of December 1987) with a variety of investment styles that rely (at least partially) on equity holdings. Note that few, if any, managers from the sample are explicit timers. We single out timers from the sample in three different ways. First, we focus on those funds from the sample for which Morningstar reported Asset Allocation as their prospectus objective. Second, we utilize a simple form of Sharpe style analysis (Sharpe, 1992) to identify implicit timers, i.e., those who exhibit the greatest variation of non-negative implied portfolio weights allocated to the market (proxied by S&P 500 index total returns) and the riskless asset (proxied by 30-day T-Bill total return). Third, we appeal to the classification provided by Brown and Goetzmann (1997), wherein one of the categories of mutual funds, the so-called "Glamour" category, was found to possess characteristics of timing (Brown and Goetzmann, 1997, p. 390). A total of 43 funds from our overall sample of 558 funds were classified by Brown and Goetzmann (1997) as "Glamour" funds.[6]

[5]Requiring instead that the Morningstar Category of the fund should *not* be Specialty (Precious Metals, Natural Resources, Technology, Utilities, Health, Financial, Real Estate, Communication, Unaligned, Convertibles) identifies the same funds.

[6]The Brown and Goetzmann classification of mutual funds into eight categories as specified in Brown and Goetzmann (1997) is available at `http://viking.som.yale.edu`.

Our results may be affected by survivorship bias because Morningstar does not report any data on disappearing funds. Several studies have generally found that survivorship may affect performance studies.[7] Within the context of this paper, survivorship bias may be of some concern because certain critical events (e.g., the crash of October 19, 1987) may have caused timing funds to either disappear or survive. The performance of the surviving timing funds may be particularly (upward) biased. To address this issue, we compare the estimates of timing and selection skill obtained for our sample of 558 funds to the estimates of mutual funds that meet the same criteria, but have disappeared prior to the end of 1996. The source of the latter data is the 1996 CRSP Survival Bias Free Mutual Fund Database.

Total monthly return, income return, and capital appreciation on the S&P 500 index, total monthly return on 30-day Treasury bills, monthly bond default premium, monthly bond horizon premium, and various stock and bond indexes were obtained from Ibbotson Associates. Total daily return and capital appreciation on the S&P 500 index were obtained from DataStream. Monthly returns on the SML and HMB factors were generously provided by Ken French.

IV.3. Results

For each fund from the sample we estimate Equations 1 through 4 using OLS regression. Standard errors are corrected for heteroskedasticity and autocorrelation of disturbances by the Newey and West (1987) correction procedure (with up to three lags). All returns are expressed in percent per month.

IV.3.A. Tests based on the CAPM

Panel A in Table 3 displays the results of estimation of the classic HM monthly test (Equation 1) summarized by Morningstar categories. One hundred ninety-seven funds feature a positive timing coefficient γ. However, only 16 of the positive timing coefficients are statistically significant at the standard 5% significance level. Large Blend and Large Growth funds each featured more than one-half of the funds from their respective universes with a positive timing coefficient; for both categories about 7.5% of all the funds featured a statistically significant positive coefficient (ten out of 132 and 4 out of 52, respectively). Notably, Mid-Cap funds, Small funds, and Domestic Hybrid funds did not boast considerable percentages of successful market timers. Two hundred ninety-seven funds feature a positive selection

[7] See, e.g., Elton, Gruber, and Blake (1996b), Brown, Goetzmann, Ibbotson, and Ross (1992), and Brown, Goetzmann, and Ross (1995).

coefficient α, out of which 32 were statistically significant at the standard 5% significance level. Furthermore, in almost all categories (with the exception of Mid-Cap Blend funds) average alphas and gammas have the opposite signs, which is consistent with the findings reported by several earlier studies of market timing.[8]

Table 3 about here

Panel B in Table 3 displays results of the adjusted timing test. Interestingly, only 109 funds feature a positive timing coefficient γ. Amazingly, only two funds had positive timing coefficients that were also statistically significant. While Large Value funds, Large Blend funds, and Domestic Hybrid funds had roughly a quarter to one third of funds in each category with positive timing coefficients, the number of statistically significant ones at the standard 5% level paled into insignificance for each of the categories. On the other hand, as many as 420 funds had a positive selection coefficient α, out of which 60 were statistically significant at the standard 5% significance level. Finally, average alphas and gammas have opposite signs for all ten Morningstar categories featured in Panel B, which suggests that there is a negative correlation between the two even under the adjusted measure.

A comparison of Panels A and B reveals that there is a consistent pattern of change for average alphas across categories as the estimation changes from the classic monthly HM tests to the adjusted test introduced herein. Average alphas increase for all ten Morningstar categories. The changes range between 8 basis points and 120 basis points. Interestingly, when size is controlled for, Value funds consistently feature the smallest increase in average alpha, while Growth funds feature the largest; when investment style is controlled for, Large funds feature the smallest increase in average alpha, while Small funds feature the largest. Respective average gammas for each category change as well, but the direction and magnitude of the change differ. The largest changes in the average gammas are a 0.145 drop for Large Growth funds, a 0.125 increase for Small Value funds, and a 0.235 increase for Small Blend funds. The differences among average alphas and average gammas under the two measures move in tandem for Value funds, Small funds, and Domestic Hybrids, i.e., for 6 out of 10 Morningstar categories; differences move in opposite directions for the remaining four categories.

IV.3.B. Tests based on the Fama-French 3-Factor Model

Panel A in Table 4 displays the results of estimation of the HM-FF3 monthly test (Equation 3) summarized by Morningstar categories. As many as 345 funds feature a positive timing

[8] See, e.g., Kon (1983), Henriksson (1984), and Jagannathan and Korajczyk (1986).

coefficient γ, out of which 31 are statistically significant at the standard 5% significance level. In fact, with the exception of Small Blend and Small Growth categories, in our sample of funds each category featured roughly one-half or more of its funds with a positive timing coefficient. Two hundred and six funds feature a positive selection coefficient α, out of which only 14 were statistically significant at the standard 5% significance level. Furthermore, in all categories average alphas and gammas have the opposite signs.

Table 4 about here

Panel B in Table 4 displays results of the adjusted-FF3 timing test (Equation 4). Two-hundred sixty eight funds feature a positive timing coefficient γ, out of which only 12 are statistically significant. With the exception of Small Blend funds and Small Growth funds, each category of funds from our sample had roughly forty percent or more of its funds with positive timing coefficients. Finally, the average alphas and gammas have opposite signs for the majority of Morningstar categories featured in Table 4 (the exception are Large Value funds, Large Blend funds, Mid-Cap Value funds, and Domestic Hybrid funds), which suggests that there is a negative correlation between the two even under the adjusted measure.

A comparison of Panels A and B in Table 4 parallels the earlier comparison of Panels A and B in Table 3, albeit on a lesser scale. There is again a consistent pattern of change for average alphas across categories as the estimation changes from the HM-FF3 tests to the adjusted-FF3 test, albeit on a considerably smaller scale. Average alphas increase for all the ten Morningstar categories, but this time the magnitude of the change is at most 28 basis points. When size is controlled for, Value funds still feature the smallest increase in average alpha, while Growth funds still feature the largest; however, these effects are negligible. Unlike the previous comparison, when investment style is controlled for, any differences between the change in average alphas between, e.g., Large funds and Small funds all but disappear — they are consistently within several basis points. Respective average gammas for each category change as well, but the magnitude of the change is again much smaller than before. The largest changes in the average gammas are increases for Mid-Cap Blend funds and Small funds, and are each up to about 0.09 in magnitude. Finally, and unlike the results obtained for CAPM-based measures, the differences among average alphas and average gammas under the two measures move in opposite directions for all ten categories of funds.

IV.3.C. Discussion

It appears that the four tests present somewhat different pictures about the selection and timing abilities of the mutual fund managers from our sample. Additional insights can be obtained by focusing on the cross-sectional distributions of the 558 selection and timing coefficients under the HM monthly specification and the adjusted specification (displayed in Figure 1), and under the HM-FF3 specification and the adjusted-FF3 specification (displayed in Figure 2). The relevant statistics are summarized in Table 5.

Figure 1 about here

Figure 2 about here

Table 5 about here

Both adjusted timing measures feature much "tighter" cross-sectional distributions of adjusted timing coefficients than either of the non-adjusted timing measures do. Indeed, Table 5 readily reveals that the cross-sectional standard deviation of each adjusted timing measure is at least two times smaller than that of the corresponding non-adjusted timing measure (0.0963 *vs.* 0.2039 and 0.0635 *vs.* 0.1611, respectively). Of course, the primary objective of conducting these tests is to determine whether manager possess statistically significant positive timing ability. Consequently, the distribution of point estimates is less important than the distribution of their t-statistics. Moreover, as was noticed by Merton (1981), the value of forecasts generated by perfect timers increases if the number of forecasts per period increases. In that sense, it is appropriate that the estimates of gamma under adjusted specification be smaller — the underlying value added by daily timing is greater than the value added by monthly timing.[9]

Each of the four specifications sheds a different light on the mutual funds from our sample. Generally, adjusted measures are more favorably disposed towards finding selection skill than their non-adjusted counterparts are. The situation with identifying timing skill is exactly the opposite. Furthermore, moving from CAPM-based measures to the measures based on

[9] Merton (1981) showed that, under standard assumptions, the value added by a perfect timer who forecasts n times per period exceeds the value added by a perfect timer who forecasts only once per period by a factor of $(2^n\Phi^n(\frac{1}{2}\sigma\sqrt{T/n})-1)/(2\Phi(\frac{1}{2}\sigma\sqrt{T})-1)$, where σ is the (constant) variance rate which prevails during the forecast period, T is the forecast period, and $\Phi(\cdot)$ denotes the cumulative normal density function (Merton, 1981, p. 374-375). For a realistic value of $\sigma = 0.2$, one-month period $T = 1/12$, and $n = 21$ forecasts of the daily timer, the value of this fraction is approximately 4.82.

the Fama-French 3-factor model brings a decrease in the assessment of selection skill and an increase in the assessment of timing skill.

A pivotal question at this point is which of the four specifications is the least biased. In Section 2 we presented simulations results which indicate that adjusted measures have more power to detect timing skill than non-adjusted measures do. Tables 3 and 4 already offer an indication that the specifications based on the Fama-French 3-factor model may be superior to those based on CAPM. For example, the classic HM specification (Equation 1) would lead to the conclusion that Small funds exhibit considerably better selection skill on average than any other category. At the same time, according to Panel A in Table 3, Small funds are plagued with negative timing skill to the extent unseen in other categories. While it is certainly possible that there may be systematic asymmetry in manager talent (for both selection and timing) across categories, a far more plausible explanation is that the anomaly detected in Panel A of Table 3 is in no small part due to the fact that Small funds hold mostly small stocks in their portfolios, and risk-return characteristics of small stocks are a well-known CAPM anomaly (Banz, 1981; Fama and French, 1992). Indeed, Panel A in Table 4 indicates that the Fama-French 3-factor model mitigates the impact of stock size on selection and timing measures.[10]

In order to explore whether this intuition is correct, we will follow an approach similar to those employed by Kothari and Warner (1997) and Ferson and Schadt (1996). Namely, both papers employ a simple yet powerful argument that naïve methods of portfolio selection should not exhibit any selection skill nor any timing skill.[11] Kothari and Warner (1997) define naïve strategies that pick stocks at random and change the portfolio periodically (each time picking stocks at random) at a rate that is consistent with turnover rates of typical mutual funds. Ferson and Schadt (1996) define three naïve strategies of investing into broad asset classes (large stocks, small stocks, government bonds, and low-grade bonds). They define an initial asset mix (65/13/20/2) and simulate three strategies: buy-and-hold, monthly rebalancing, and annual rebalancing.[12]

We do not expect any of the four tests to be unbiased and efficient. After all, Kothari and Warner (1997) have already indicated that the classic HM test and the HM-FF3 test, i.e., the non-adjusted tests (Equations 1 and 3) may uncover abnormal performance for naïve

[10]Analogous conclusions can be reached for the adjusted measures by comparing the results displayed in Panel B of Table 3 to those from Panel B in Table 4.

[11]An interesting question in its own right is what constitutes a naïve strategy and where the line between naïve and not-so-naïve strategies should be drawn.

[12]Both rebalancing strategies rebalance to the same asset mix – 65/13/20/2.

strategies that employ neither selection nor timing skill. Instead, we pursue a pragmatic task of identifying the specification that appears to be the least biased among the four. To that end, we apply the four tests on managed funds which did not appear to exert timing effort and compare the results.

We focus on two varieties of non-timing funds. The first variety includes the six index funds from our sample; the second variety consists of 55 fictitious index funds that would track various stock indexes which are available for the period from January 1988 to March 1998 from Ibbotson Associates. Naturally, we should expect these two varieties of funds to exhibit neither selection nor timing (in)ability.

Table 6 about here

Table 6 displays results of running the four tests on the six index funds we identified in our sample. A comparison of Panels A and B (based on the CAPM) to Panels C and D (based on the Fama-French 3-factor model) suggests that, as expected, the specifications based on the Fama-French 3-factor model produce estimates of timing skill that are less biased than those based on the CAPM. Specifically, Panels C and D each featured one fund with a statistically significant timing coefficient gamma at the 10% level (or less), whereas Panels A and B each featured three such funds. At the same time, Panels C and D of Table 6 do not seem to feature a clear distinction between the results obtained from the HM-FF3 specification and those obtained from the adjusted-FF3 specification. A similar pattern exists for alphas. It is, of course, very difficult to draw any conclusions on the basis of only six funds. To overcome this limitation, we turn our attention to the 55 stock indexes.

Table 7 about here

Table 7 displays the summary statistics of the results of running the four tests on the 55 stock indexes. It reaffirms the finding we discussed above: the specifications based on the Fama-French 3-factor model are less biased than those based on CAPM, as witnessed both by the properties of the cross-sectional distributions of the respective alphas and gammas and by the number of respective alphas and gammas that are different from 0 at standard significance levels. Furthermore, Table 7 indicates that there is a trade-off between the precision with which the two Fama-French 3-factor-based specifications measure alpha and gamma: a larger standard deviation of alpha measured by the adjusted-FF3 specification (0.1829 *vs.* 0.1332) is compensated for by a smaller standard deviation of gamma (0.0251 *vs.* 0.0767). Nevertheless, the number of statistically significant non-zero alphas and gammas (at the 10% significance level) is consistently smaller for the adjusted-FF3 specification than for the HM-FF3 specification (11 *vs.* 20 and 6 *vs.* 11, respectively). We thus conclude that the adjusted-FF3 specification is slightly less biased than the HM-FF3 specification.

IV.3.D. How Did the Timers Perform?

We do not have the information whether some of the funds from our sample are timing the market. It is quite likely that most of the funds from the sample do not behave like (nor profess to be) market timers. It would be particularly interesting to see whether those managers who are likely to be timers exhibit timing skill. We attempt to identify timers in our sample in three different ways. First, we identify the funds from our sample that had stated Asset Allocation as their Prospectus Objective. According to Morningstar's on-line description, managers of asset allocation funds "... often use a flexible combination of stocks, bonds, and cash; some, but not all, shift assets frequently based on analysis of business-cycle trends." Second, we conduct a simple form of Sharpe's style analysis (Sharpe, 1992). For each fund we compute Sharpe weights for two asset classes, the S & P 500 and the 30-day Treasury Bill, using a 12-month rolling window. We used the volatility of the resulting weight[13] on the S & P 500, defined here as the sum of absolute values of successive period weight changes, as an (admittedly noisy) proxy for the intensity of market timing efforts. Third, it is possible that Morningstar classification does not properly identify market timers. To that end, we employ a different classification of funds proposed by Brown and Goetzmann (1997). We extract from our sample the 43 funds that were classified as "Glamour" funds, i.e., funds that may engage in some timing activity, by Brown and Goetzmann (1997) and report summary statistics of timing tests.

We identified 23 Asset Allocation funds in our sample. Not surprisingly, all but three belong to the Domestic Hybrid category.[14] For this reason, we need to also consider an asset pricing model that would incorporate bonds. We thus add excess returns on CS First Boston's High Yield Corporate Bond Index and on Ibbotson's Long Term Government Bond Index to the three Fama-French factors.[15] Both indexes are available from Ibbotson Associates. We

[13]Note that the weight on the 30-day Treasury Bill and the weight on the S & P 500 sum to 1 (i.e., to 100%) by construction.

[14]The remaining three funds belong to Large Value, Large Blend, and Small Value categories, respectively.

[15]A similar approach was utilized by Elton, Gruber, and Blake (1996a). They use one bond index defined as a par-weighted combination of the Lehman Brothers Aggregate Bond Index and the Blume/Keim High-Yield Bond Index. Also, our approach is similar to that advocated by Fama and French (1993); our two excess returns on bond indexes capture the term premium and the default premium.

call the resulting non-adjusted 5-factor model **HM-5** and define it as follows:

$$Z_{p,t} = \alpha + \beta_1 Z_{m,t} + \gamma \max\{-Z_{m,t}, 0\} + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \beta_4 \mathrm{HiYldCorp}_t + \beta_5 \mathrm{USLTGvt}_t + \epsilon_t. \quad (5)$$

Similarly, we call the resulting adjusted 5-factor model **adjusted-5** and define it as follows:

$$Z_{p,t} = \alpha + \beta_1 Z_{m,t} + \gamma P_{m,t} + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \beta_4 \mathrm{HiYldCorp}_t + \beta_5 \mathrm{USLTGvt}_t + \epsilon_t. \quad (6)$$

The results of all six tests are summarized in Table 8. While each of the six tests produced a certain number of positive timing coefficients, only the two 5-factor tests found a small number of funds with positive timing coefficients that were statistically significant at the standard 5% level; the HM-5 test uncovered two such funds, while the adjusted-5 test found only one. The overall conclusion is that very few of the Asset Allocation funds from our sample displayed timing skill.

Table 8 about here

Our second attempt at identifying timers is based on the volatility of implied asset weights in the manner described above. We identified the funds that were in the top 10 and in the top 5 percentile, as well as in the bottom 10 and in the bottom 5 percentile of all the funds in our sample with respect to the volatility of implied asset weights. It turned out that most of the funds from the top 10 percentile were Small Value funds (twenty-two out of 56); MidCap and Small funds together account for the vast majority of the funds (50 out of 56). On the other hand, forty-five out of 55 funds from the bottom 10 percentile were Large funds; Mid-Cap funds accounted for 9 out of the remaining 10 funds, and one fund was a Small Growth fund. The first interesting observation is that neither the top 10 percentile funds nor the bottom 10 percentile funds featured a substantial presence of Domestic Hybrid funds, i.e., the vast majority of funds under present consideration were primarily stock funds. In view of this observation, we did not perform the tests based on 5-factor specifications. The second interesting observation is at the same time a *caveat*; the fact that most of the bottom/top 10 percentile funds were Large/Small funds may indicate that the volatility of implied asset weights may be an artifact of small stock phenomena as well as (or perhaps even rather than) timing.

Table 9 about here

The results of the four tests are summarized in Table 9. The classic HM test and the adjusted test find little timing skill in the top 10 percentile (one and zero statistically significant positive timing coefficients, respectively). In fact, according to both CAPM based measures, the bottom 10 percentile exhibited more timing skill than did the top 10 percentile! The two tests based on the Fama-French 3-factor model both found slightly more positive timing skill among the funds from the top 10 percentile than among the funds from the bottom 10 percentile. The results of the HM-FF3 test indicate that the top 10 percentile featured 7 statistically significant positive timing coefficients, whereas the bottom 10 percentile featured 5. Similarly, the results of the adjusted-FF3 test indicate that the top 10 percentile featured 3 statistically significant positive timing coefficients, whereas the bottom 10 percentile featured 2. We conclude that the top 10 percentile did not considerably outperform the bottom 10 percentile with respect to timing skill.[16] In sum, the funds that are likely to be involved in timing activities for the most part did not demonstrate superior timing skill.

Finally, our third attempt at identifying timers consisted of identifying "Glamour" funds (Brown and Goetzmann, 1997) from our sample. Brown and Goetzmann (1997) proposed a classification of mutual funds into eight categories and have demonstrated that such a classification is superior to the widely used Morningstar classification. One of the categories identified by Brown and Goetzmann, the "Glamour" category, was characterized as "... domestic 'trend-chasers', displaying positive correlation to preceding S&P index returns" (Brown and Goetzmann, 1997, p. 390). Our sample contained 43 "Glamour" funds. Results of timing tests performed on the 43 "Glamour" funds are summarized in Table 10. Table 10 suggests that, similarly to the above two approaches to identifying timers, very few of the "Glamour" funds from our sample displayed timing skill.

Table 10 about here

IV.3.E. Effect of Survivorship

Our final analysis is aimed at documenting the effect of survivorship on our results. We compare the estimates of timing and selection skill obtained for our sample of 558 funds that have survived in the period from January 1988 to March 1998 to the estimates of mutual funds that existed in January 1988, but have since become defunct. Specifically, we look into

[16] The same conclusion can be reached on the basis of a comparison between the top 5 percentile and the bottom 5 percentile.

horizons of two to seven years and for each horizon of k years, $k = 2, \ldots, 7$ years compare the performance of our surviving 558 funds to all the non-surviving funds that have survived for at least k years.[17] The non-surviving funds were obtained from the 1996 CRSP Survival Bias Free Mutual Fund Database. The selection criteria was matched as closely as possible to those employed to extract the 558 surviving funds from the April 1998 Morningstar Principia CD-ROM, except that the fund had to be "Dead" by the end of 1996 and, at the same time, had to survive for at least two years from January 1988 (so as to allow for a sufficient length of the time series of returns). The results of performing the timing tests on both surviving and non-surviving funds for horizons from two to seven years are presented in Table 11.

Table 11 about here

Table 11 suggests that, for virtually every horizon k and every test, surviving funds exhibited a larger average alpha than did their non-surviving counterparts (the magnitude of the difference ranged from about 10 basis points to 40 basis points per month). As was the case in Table 3, Panel B and in Table 5, the average alpha resulting from the adjusted test (Equation 2) was typically larger by 30-80 basis points than the average alpha resulting from any of the remaining five tests, both for surviving and non-surviving funds. Furthermore, it was also typical that larger percentages of surviving funds had a statistically significant positive alpha, particularly at longer horizons. The estimates of timing skill paint a somewhat more controversial picture. That is, average gammas for surviving funds were typically higher than the average gammas for non-surviving funds for shorter horizons ($k = 2, 3$), and were typically lower for longer horizons ($k = 4, \ldots, 7$). On the other hand, the percentages of statistically significant positive gammas were similar for most horizons.

We also provided for a t-test of statistical significance of the difference of cross-sectional means of alphas and gammas for each specification and each of the horizons (the last two columns in Table 11). It should be noted that the test is intended for illustrative purposes only; it is a standard t-test, the implementation of which does not take into account cross-sectional correlation between the point estimates of individual alphas and gammas, respectively.[18] With the exception of t-statistics reported for very short horizons, i.e., $k = 2, 3$ (which are based on very short time series of data), and with the exception of t-statistics

[17]We limit our analysis to seven years because the number of non-surviving funds that survived at least seven years is only 31, and looking at the eight-year horizon would limit the analysis even further to only 16 non-surviving funds.

[18]See Elton, Gruber, and Blake (1996b, p. 1107, fn. 16) for an outline of a more elaborate procedure which takes such cross-sectional correlations into account.

reported for tests of timing skill based on Equation 1 (which have been shown earlier in this paper to be fairly biased), all of the t-statistics pertaining to alphas exhibit values that are positive and statistically significant at standard levels, often with values of 4 or more, whereas the absolute value of none of the (typically negative) t-statistics pertaining to gammas exceeds 1.4.

The sign and magnitude of t-statistics reported for gammas together indicate that the basic finding, which suggests that there are no significant differences in the cross-sectional means of gammas for a variety of specifications and for longer horizons, would almost certainly persist under a more elaborate statistical procedure that takes the aforementioned cross-sectional correlations among estimates of gammas into consideration. On the other hand, the typical magnitude of t-statistics reported for the difference of cross-sectional means of alphas indicates that statistically significant differences would very likely persist under such a statistical procedure.

In sum, it appears that the primary distinguishing factor between surviving and non-surviving funds is their alpha. It also appears that neither the magnitude nor the percentage of statistically significant positive gammas (i.e., quantities of primary interest in this study) in the cross-section of funds were strongly affected by survivorship bias.

V. Conclusion

Simulations of market timing strategies under reasonable assumptions indicate that the widely used Henriksson-Merton parametric test has low power to detect timing skill. In addition, the information about the value of the implied put given by the regression coefficients is strongly biased downwards. The reason for the failure of the HM monthly test statistics of timing in our simulation experiment is simply that the market timer makes decisions at a more frequent interval than the one over which the value of the implied put is calculated.

We propose an adjusted test of timing skill, i.e., a simple correction for the above problem. While our simulations suggest that adjusted tests are not as powerful as HM tests performed directly on daily timer data, the adjusted approach has the advantage of not requiring daily timer data to be obtained. Instead we rely upon an instrument developed from the daily returns to an index correlated to the timer's risky asset. We find that the adjusted test of timing skill has some power to detect even moderate timing skill.

In our empirical analysis we explored the effect of the proposed adjusted measure on inferences about market timing skill. This required us, *inter alia*, to address the effects of passive timing due to the choice of investment style.

We focused on four tests of timing skill: the classic HM test, the adjusted test, the HM-FF3 test, and the adjusted-FF3 test. Very few funds from our sample of 558 mutual funds exhibited statistically significant positive timing skill under either measure.

We find that the adjusted-FF3 test, a modification of the classic HM test adjusted for daily frequency of timing activities and for exposure to the three risk factors identified by Fama and French (1992, 1993, 1996), presents a relatively unbiased approach to measuring timing and selectivity in mutual fund returns. Applying these same measures to known passively managed portfolios shows the extent to which investment style affects inference about timing skill.

References

[1] Admati, A., S. Bhattacharya, P. Pfleiderer, and S. Ross, 1986. On Timing and Selectivity. *The Journal of Finance*, **41**(3), 715-732.

[2] Admati, A. and S. Ross, 1985. Measuring Investment Performance in a Rational Expectations Equilibrium Model. *Journal of Business*, **58**(1), 1-26.

[3] Banz, R., 1981. The Relationship Between Return and Market Value of Common Stocks. *Journal of Financial Economics*, **9**, 3-18.

[4] Blazenko, G., P. Boyle, and K. Newport, 1990. Valuation of Tandem Options. In Fabozzi, F. (Ed.), Advances in Futures and Options Research: A Research Annual, *Volume 4*, Jai Press, Greenwich, CT, London, England, 39-49.

[5] Brocanto, J. and P. Chandy, 1994. Does Market Timing Really Work in the Real World? *Journal of Portfolio Management*, **20**(3), 39-44.

[6] Brown, S., W. Goetzmann, and A. Kumar, 1997. The Dow Theory: William Peter Hamilton's Track Record Re-Considered. *The Journal of Finance*, forthcoming.

[7] Brown, S., W. Goetzmann, R. Ibbotson, and S. Ross, 1992. Survivorship Bias in Performance Studies. *The Review of Financial Studies*, **5**(4), 553-580.

[8] Brown, S., W. Goetzmann, and S. Ross, 1992. Survival. *The Journal of Finance*, **50**(3), 853-873.

[9] Busse, J., 1997. Volatility Timing in Mutual Funds: Evidence from Daily Returns. Working Paper, Stern School of Management, NYU.

[10] Chance, D. and M. Hemler, 1998. The Performance of Professional Timers: Evidence from Executing Timing Strategies. Working Paper, Pamplin College of Business, Virginia Tech.

[11] Chang, E. and W. Lewellen, 1984. Market Timing and Mutual Fund Performance. *Journal of Business*, **57**, 57-72.

[12] Chen, Z. and P. Knez, 1996. Portfolio Performance Measurement: Theory and Applications. *The Review of Financial Studies*, **9**, 511-555.

[13] Cumby, R. and J. Glen, 1990. Evaluating the Performance of International Mutual Funds. *The Journal of Finance*, **45**(2), 497-521.

[14] Cumby, R. and D. Modest, 1987. Tests for Market Timing Ability: A Framework for Forecast Evaluation. *Journal of Financial Economics*, **19**, 169-189.

[15] Dybvig, P. and S. Ross, 1985. Differential Information and Performance Measurement Using a Security Market Line. *The Journal of Finance*, **40**(2), 383-399.

[16] Elton, E., M. Gruber, and C. Blake, 1996a. The Persistence of Risk-Adjusted Mutual Fund Performance. *Journal of Business*, **69**(2), 133-157.

[17] Elton, E., M. Gruber, and C. Blake, 1996b. Survivorship Bias and Mutual Fund Performance. *The Review of Financial Studies*, **9**(4), 1097-1120.

[18] Fama, E. and K. French, 1992. The Cross-Section of Expected Stock Returns. *The Journal of Finance*, **47**(2), 427-465.

[19] Fama, E. and K. French, 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, 3-56.

[20] Fama, E. and K. French, 1996. Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, **51**(1), 55-84.

[21] Ferson, W. and R. Schadt, 1996. Measuring Fund Strategy and Performance in Changing Economic Conditions. *The Journal of Finance*, **51**(2), 425-461.

[22] Glosten, L. and R. Jagannathan, 1994. A Contingent Claims Approach to Performance Evaluation. *Journal of Empirical Finance*, **1**, 133-160.

[23] Graham, J. and C. Harvey, 1996. Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendation. *Journal of Financial Economics*, **42**, 397-421.

[24] Grinblatt, M. and S. Titman, 1989a. Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings. *Journal of Business*, **62**(3), 393-416.

[25] Grinblatt, M. and S. Titman, 1989b. Portfolio Performance Evaluation: Old Issues and New Insights. *The Review of Financial Studies*, **2**(3), 393-421.

[26] Grinblatt, M. and S. Titman, 1993. Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns. *Journal of Business*, **66**(1), 47-68.

[27] Grinblatt, M. and S. Titman, 1994. A Study o Monthly Mutual Fund Returns and Performance Evaluation Techniques. *Journal of Financial and Quantitative Analysis*, **29**(3), 419-444.

[28] Grinblatt, M. and S. Titman, 1995. Performance Evaluation. In Jarrow, R., V. Maksimovic, and W. Ziemba (Eds.), Handbook in Operations Research and Managements Science, Vol. 9: Finance, Elsevier, Amsterdam, New York, 581-609.

[29] Henriksson, R. and R. Merton, 1981. On Market Timing and Investment Performance. II. Statistical procedures for Evaluating Forecasting Skills. *Journal of Business*, **54**(4), 513-533.

[30] Henriksson, R., 1984. Market Timing and Mutual Fund Performance. An Empirical Investigation. *Journal of Business*, **57**(1), 73-96.

[31] Jagannathan, R. and R. Korajczyk, 1986. Assessing the Market Timing Performance of Managed Portfolios. *Journal of Business*, **59**(2), 217-235.

[32] Jensen, M., 1968. The Performance of Mutual Funds in the Period 1945-1964. *The Journal of Finance*, **23**(2), 389-416.

[33] Jensen, M., 1969. Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios. *Journal of Business*, **42**(April), 167-247.

[34] Kon, S., 1983. The Market-Timing Performance of Mutual Fund Managers. *Journal of Business*, **56**, 323-347.

[35] Kothari, S. and J. Warner, 1997. Evaluating Mutual Fund Performance. Working Paper, August 1997.

[36] Kraus, A. and R. Litzenberger, 1976. Skewness Preference and the Valuation of Risky Assets. *The Journal of Finance*, **31**, 1085-1100.

[37] Lehmann, B. and D. Modest, 1987. Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons. *The Journal of Finance*, **42**(2), 233-265.

[38] Lintner, J., 1965. Security Prices, Risk, and Maximal Gains from Diversification. *The Journal of Finance*, **20**(4), 587-615.

[39] Low, C., 1999. Implicit Timing and Pseudo-Options in Stock Bulls and Bears. Yale SOM Working Paper, February 1999.

[40] Mayers, D. and E. Rice, 1979. Measuring Portfolio Performance and the Empirical Content of Asset Pricing Models. *Journal of Financial Economics*, **7**, 3-28.

[41] Merton, R., 1981. On Market Timing and Investment Performance. I. An Equilibrium Theory of Value for Market Forecasts. *Journal of Business*, **54**(3), 363-406.

[42] Mossin, J., 1966. Equilibrium in a Capital Asset Market. *Econometrica*, **34**(4), 768-783.

[43] Newey, W. and K. West, 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, **55**(3), 703-708.

[44] Pfleiderer, P. and S. Bhattacharya, 1983. A Note on Performance Evaluation. Technical Report No. 714, Stanford, CA.: Stanford University, Graduate School of Business.

[45] Pesaran, M. and A. Timmermann, 1994. A Generalization of the Non-Parametric Henriksson-Merton Test of Market Timing. *Economics Letters*, **44**, 1-7.

[46] Roll, R., 1978. Ambiguity When Performance is Measured by the Securities Market Line. *The Journal of Finance*, **33**(4), 1051-1069.

[47] Roll R., 1979. A Reply to Mayers and Rice (1979). *Journal of Financial Economics*, **7**, 391-400.

[48] Rosenberg, B., K. Reid, and R. Lanstein, 1985. Persuasive Evidence of Market Inefficiency. *Journal of Portfolio Management*, **11**, 9-17.

[49] Sharpe, W., 1964. Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance*, **19**(3), 425-442.

[50] Sharpe, W., 1992. Asset allocation: Management style and performance measurement. *The Journal of Portfolio Management*, **Winter 1992**, 7-19.

[51] Treynor, J., 1965. How to Rate Management of Investment Funds. *Harvard Business Review*, **43**(January-February 1965), 63-75.

[52] Treynor, J. and K. Mazuy, 1966. Can Mutual Funds Outguess the Market? *Harvard Business Review*, **44**(July-August 1966), 131-136.

[53] Wagner, J., S. Shellans, and R. Paul, 1992. Market Timing Works Where it Matters Most ... In the Real World. *Journal of Portfolio Management*, **18**(4), 86-90.

Table 1

Simulated Mean Values of Selection and Timing Coefficients

The table reports mean values of coefficients from three tests of timing skill on simulated data for a range of timing skill. Three specifications are simulated. The regression equation for "Daily HM" and "Monthly HM" has the same specification, given in Equation 1, except that the frequency of observation differs (daily *vs.* monthly). "Adjusted" test also uses monthly values for the dependent variable and the market variable, but follows Equation 2. That is, it substitutes $P_{m,t} = \left[\left(\prod_{\tau \in month(t)}^{t} \max\{1 + R_{m,\tau}, 1 + R_{f,\tau}\}\right) - 1\right] - R_{m,t}$ for the standard Henriksson-Merton term $\max\{-Z_{m,t}, 0\}$. One thousand simulations are performed for each skill level. Ten years of daily market premia are generated as i.i.d. lognormal random variables with annualized mean of 0.1 and annualized standard deviation of 0.16. Skill levels range from no foresight (*skill*=0) to perfect foresight (*skill*=1). Intermediate skill levels (*skill*=0.1, 0.2, ..., 0.9) are simulated by allowing the timer to correctly forecast the sign of the next day's excess return on the market with probability p=*skill*. For example, *skill*=0.7 indicates that the timer will correctly forecast the sign of the next day's excess return on the market 70% of the time. We simulate the timer's strategy as 100% in the market/riskless asset when the forecast excess return on the market for the day is positive/negative. The bold-faced entries in the table are inserted by definition; an actual attempt to run the specified regression for the perfect daily HM-style timer would lead to a perfect fit. Values of α are expressed in percent per month.

					α values						
skill	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Daily HM	0.0000	0.0007	0.0011	0.0010	0.0007	0.0004	0.0007	0.0003	0.0007	0.0005	**0.0000**
Monthly HM	-7.7528	-6.2423	-4.7108	-0.1676	-1.5533	0.0614	1.6820	3.3423	5.0223	6.7375	8.4659
Adjusted	-0.6392	-0.4417	-0.2913	-0.1576	-0.0898	-0.0133	0.0601	0.0627	0.0750	0.0236	0.0000
					β values						
skill	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Daily HM	0.0000	0.0988	0.1981	0.2990	0.3984	0.4993	0.5998	0.6992	0.7983	0.8993	**1.0000**
Monthly HM	0.3594	0.3833	0.4083	0.4286	0.4581	0.4850	0.5144	0.5406	0.5694	0.6008	0.6352
Adjusted	0.0676	0.1470	0.2298	0.3029	0.4062	0.4981	0.5908	0.6878	0.7983	0.8933	1.000
					γ values						
skill	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Daily HM	-1.0000	-0.8019	-0.6034	-0.4031	-0.2017	-0.0001	0.1988	0.3994	0.5978	0.7987	**1.0000**
Monthly HM	-0.1986	-0.1665	-0.1323	-0.3595	-0.0666	-0.0291	0.0147	0.0516	0.0915	0.1338	0.1808
Adjusted	-0.8499	-0.6939	-0.5296	-0.3567	-0.1801	0.0026	0.1876	0.3838	0.5815	0.7911	1.0000

Table 2

Quantiles of Critical t-value

The table reports the quantile of the critical t-value of 1.96 on the timing coefficient γ in the distribution of t-values generated by 1,000 simulations for each skill level and for each of the three specifications described in the caption to Table 1. The columns report the power of the different specifications of the test to detect skill for skill levels of *skill*=0.1, 0.2, ...,1. The column under *skill*=0.5 reports results under the null hypothesis of no timing skill. For example, the null hypothesis of no timing ability is rejected about 57% of the time for a timer with perfect foresight when the standard Henriksson-Merton specification with monthly data is used. In addition to the usual three tests, we also report the results of a Treynor-Mazuy specification for both daily and monthly data. The powers of the two models are very similar for both tests on daily and on monthly data. The bold-faced entry in the table is inserted by definition; an actual attempt to run the specified regression for the perfect daily HM-style timer would lead to a perfect fit.

Quantiles for t = 1.96											
skill	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Daily HM	1.000	1.000	1.000	1.000	1.000	0.879	0.005	0.000	0.000	0.000	**0.000**
Monthly HM	1.000	1.000	0.998	0.990	0.986	0.983	0.966	0.951	0.897	0.800	0.570
Adjusted	1.000	1.000	1.000	1.000	0.999	0.959	0.732	0.198	0.005	0.000	0.000
Daily TM	1.000	1.000	1.000	1.000	1.000	0.820	0.037	0.000	0.000	0.000	0.0000
Monthly TM	1.000	1.000	0.998	0.990	0.990	0.980	0.949	0.939	0.889	0.786	0.527

Table 3

CAPM-Based Timing Tests by Category

The table reports the results of CAPM-based timing tests (Equations 1 and 2) performed on our sample of 558 mutual funds in the period from January 1988 to March 1998 (123 monthly observations). The results are summarized by Morningstar Category classification. The funds from our sample each belong to one of the ten Morningstar Categories featured in the table. For each Morningstar Category, we report the average values of the estimated regression coefficients α, β, and γ, the corresponding average t-statistics, the average p-value, the number of positive coefficients (denoted as $t > 0$) among the funds from that category, and the number of positive coefficients across all the funds from that Morningstar Category that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). Panel A displays results of a classic HM test (Equation 1), while Panel B displays results of our adjusted test (Equation 2). The mutual fund sample and the data required to execute tests based on the CAPM are discussed in Section IV.2.

Morningstar Category (N)	Panel A: HM Tests						Panel B: Adjusted Tests					
		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
	α	0.0565	0.4278	0.3768	50	6	α	0.1689	0.3736	0.3860	52	5
Large Value ($N = 76$)	β	0.8475	14.1721	0.0000	76	76	β	0.8922	22.3428	0.0000	76	76
	γ	-0.0768	-0.7937	0.6906	22	0	γ	-0.0310	-0.5060	0.6372	20	1
	α	-0.0824	-0.5136	0.6279	48	5	α	0.1281	0.2857	0.4170	82	7
Large Blend ($N = 132$)	β	0.9186	24.9950	0.0000	132	132	β	0.9290	37.3287	0.0000	132	132
	γ	0.0048	0.0214	0.4839	68	10	γ	-0.0291	-0.5543	0.6599	33	1
	α	-0.2738	-0.8825	0.7451	10	0	α	0.2922	0.4800	0.3489	39	1
Large Growth ($N = 52$)	β	1.0937	14.9703	0.0000	52	52	β	1.0885	22.2422	0.0000	52	52
	γ	0.0805	0.5601	0.3423	36	4	γ	-0.0651	-0.8362	0.7363	9	0
	α	0.1603	0.5190	0.3560	31	5	α	0.4441	0.9229	0.2710	34	9
Mid-Cap Value ($N = 42$)	β	0.7506	10.1375	0.0000	42	42	β	0.8276	15.2543	0.0000	42	42
	γ	-0.1208	-0.6935	0.7028	8	0	γ	-0.0640	-0.9672	0.7522	8	0
	α	-0.0021	0.0809	0.4747	23	2	α	0.4347	0.8176	0.2595	33	2
Mid-Cap Blend ($N = 40$)	β	0.8515	10.4370	0.0000	40	40	β	0.8985	14.9404	0.0000	40	40
	γ	-0.0413	-0.2309	0.5797	15	1	γ	-0.0704	-0.9018	0.7657	9	0

Table 3 (continued)

CAPM-Based Timing Tests by Category

Morningstar Category (N)		Panel A: HM Tests						Panel B: Adjusted Tests				
		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
	α	0.0031	-0.0642	0.5162	27	0	α	1.0461	1.3639	0.1434	58	12
Mid-Cap Growth ($N = 61$)	β	1.0341	9.3833	0.0000	61	61	β	1.1620	14.4902	0.0000	61	61
	γ	-0.1300	-0.4659	0.6208	20	0	γ	-0.1741	-1.7412	0.9111	1	0
	α	0.4066	1.0800	0.2034	26	4	α	1.0137	1.3599	0.1244	29	4
Small Value ($N = 30$)	β	0.6389	5.7486	0.0002	30	30	β	0.8064	9.0021	0.0000	30	30
	γ	-0.2639	-0.9921	0.7984	3	0	γ	-0.1380	-1.2167	0.8556	3	0
	α	0.4058	1.2249	0.1843	14	3	α	1.0661	1.5343	0.1024	15	4
Small Blend ($N = 16$)	β	0.7039	5.7878	0.0008	16	16	β	0.9464	8.8487	0.0009	16	16
	γ	-0.4094	-1.4476	0.9026	0	0	γ	-0.1739	-1.5456	0.9201	0	0
	α	0.3259	0.5755	0.3263	23	0	α	1.5272	1.7953	0.0661	32	11
Small Growth ($N = 32$)	β	0.9402	6.8846	0.0033	32	31	β	1.1526	10.3662	0.0000	32	32
	γ	-0.2815	-0.9298	0.7766	2	0	γ	-0.2262	-1.9604	0.9522	0	0
	α	0.0679	0.4059	0.4050	45	7	α	0.1436	0.3771	0.4043	46	5
Domestic Hybrid ($N = 77$)	β	0.5518	11.7778	0.0020	77	76	β	0.5855	19.3344	0.0002	77	77
	γ	-0.0589	-0.7110	0.6646	23	1	γ	-0.0223	-0.4786	0.6184	26	0

Table 4
Fama-French 3-Factor-Based Timing Tests by Category

The table reports the results of Fama-French 3-factor-based timing tests (Equations 3 and 4) performed on our sample of 558 mutual funds in the period from January 1988 to March 1998 (123 monthly observations). The results are summarized by Morningstar Category classification. The funds from our sample each belong to one of the ten Morningstar Categories featured in the table. For each Morningstar Category, we report the average values of the five estimated regression coefficients (α, β_1, γ, β_2, and β_3), the corresponding average t-statistics, the average p-value, the number of positive coefficients (denoted as $t > 0$) among the funds from that category, and the number of positive coefficients across all the funds from that Morningstar Category that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). Panel A displays results of the HM-FF3 test (Equation 3), while Panel B displays results of our adjusted-FF3 test (Equation 4). The mutual fund sample and the data required to execute tests based on the Fama-French 3-factor model are discussed in Section IV.2.

Morningstar Category (N)	Panel A: HM-FF3 Tests						Panel B: Adjusted-FF3 Tests					
	Coefficient		t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$	Coefficient		t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
Large Value ($N = 76$)	α	-0.1900	-0.9481	0.7419	15	0	α	-0.1017	-0.3903	0.60406	25	1
	β_1	0.9500	16.3130	0.0000	76	76	β_1	0.9283	26.5343	0.0000	76	76
	γ	0.0509	0.3951	0.3851	54	6	γ	-0.0023	0.0205	0.4969	37	2
	β_2	0.2012	3.9972	0.0787	71	59	β_2	0.1969	3.9258	0.0768	70	60
	β_3	0.1620	3.2291	0.1328	68	49	β_3	0.1559	3.3323	0.1453	67	49
Large Blend ($N = 132$)	α	-0.1160	-0.7295	0.6848	36	4	α	-0.0279	-0.1641	0.5507	55	3
	β_1	0.9373	23.1458	0.0000	132	132	β_1	0.9198	38.2483	0.0000	132	132
	γ	0.0430	0.4084	0.3861	87	7	γ	-0.0038	-0.0472	0.5147	58	1
	β_2	0.1578	3.2485	0.1615	112	84	β_2	0.1539	3.1513	0.1672	112	82
	β_3	-0.0289	-0.3666	0.5616	60	23	β_3	-0.0341	-0.4869	0.5707	61	23
Large Growth ($N = 52$)	α	-0.0907	-0.3009	0.5630	24	0	α	0.0882	0.1525	0.4500	33	2
	β_1	1.0338	14.4873	0.0000	52	52	β_1	1.0147	21.1711	0.0000	52	52
	γ	0.0560	0.4323	0.3828	33	3	γ	-0.0141	-0.1464	0.5479	23	2
	β_2	0.2909	4.1686	0.0527	50	45	β_2	0.2842	4.0286	0.0557	49	45
	β_3	-0.2924	-3.9737	0.9509	2	0	β_3	-0.2995	-4.2034	0.9555	2	0
Mid-Cap Value ($N = 42$)	α	-0.1414	-0.7764	0.6744	13	0	α	-0.0083	-0.0474	0.5084	21	3
	β_1	0.8807	12.4847	0.0000	42	42	β_1	0.8591	19.9824	0.0000	42	42
	γ	0.0559	0.5213	0.3528	32	2	γ	-0.0008	-0.1665	0.5421	19	1
	β_2	0.3735	6.0526	0.0224	41	41	β_2	0.3678	5.9803	0.0221	41	41
	β_3	0.1486	2.1615	0.1518	37	23	β_3	0.1417	2.1752	0.1591	37	23
Mid-Cap Blend ($N = 40$)	α	-0.1284	-0.4698	0.6202	14	0	α	-0.0419	-0.1620	0.5447	18	0
	β_1	0.9173	12.6670	0.0000	40	40	β_1	0.8793	18.8301	0.0000	40	40
	γ	0.0818	0.5947	0.3271	31	3	γ	0.0042	0.1714	0.4493	22	0
	β_2	0.4638	8.0112	0.0101	40	39	β_2	0.4582	7.8958	0.0116	40	39
	β_3	-0.0579	-0.6124	0.5953	17	6	β_3	-0.0675	-0.7437	0.6151	14	6

Table 4 (continued)
Fama-French 3-Factor-Based Timing Tests by Category

Morningstar Category (N)		Panel A: HM-FF3 Tests						Panel B: Adjusted-FF3 Tests				
		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
Mid-Cap Growth ($N = 61$)	α	0.1120	0.2513	0.4347	35	3	α	0.3986	0.6137	0.3198	48	4
	β_1	1.0186	10.0094	0.0000	61	61	β_1	1.0653	16.8849	0.0000	61	61
	γ	-0.0568	-0.2130	0.5365	28	2	γ	-0.0523	-0.5880	0.6715	48	4
	β_2	0.7216	9.7162	0.0000	61	61	β_2	0.7174	16.8849	0.0001	61	61
	β_3	-0.3883	-4.1271	0.9618	2	0	β_3	-0.3831	-4.3191	0.9668	1	0
Small Value ($N = 30$)	α	-0.1553	-0.5447	0.6406	11	0	α	0.0056	0.0507	0.4686	16	0
	β_1	0.8876	11.7589	0.0000	30	30	β_1	0.8484	17.4227	0.0000	30	30
	γ	0.0924	0.6999	0.2975	25	4	γ	-0.0042	-0.0640	0.5443	14	2
	β_2	0.8656	14.4490	0.0000	30	30	β_2	0.8577	14.0502	0.0000	30	30
	β_3	0.2103	3.0702	0.0704	28	22	β_3	0.1992	3.0633	0.0740	28	21
Small Blend ($N = 16$)	α	0.0415	0.6498	0.3308	12	3	α	0.0758	0.3478	0.3822	11	0
	β_1	0.8773	10.8598	0.0000	16	16	β_1	0.9452	18.5741	0.0000	16	16
	γ	-0.1250	-0.7049	0.6958	4	0	γ	-0.0302	-0.3371	0.6079	5	0
	β_2	0.8951	15.2143	0.0000	16	16	β_2	0.8994	15.0461	0.0000	16	16
	β_3	0.0058	0.0078	0.5458	6	3	β_3	0.0197	0.1378	0.5271	6	3
Small Growth ($N = 32$)	α	0.3682	0.8911	0.2588	26	4	α	0.5446	0.8702	0.2524	29	3
	β_1	0.9621	8.3630	0.0033	32	31	β_1	1.0418	14.4939	0.0005	32	32
	γ	-0.1317	-0.6425	0.6909	8	1	γ	-0.0518	-0.5759	0.6774	7	0
	β_2	1.0387	12.0528	0.0000	32	32	β_2	1.0404	12.1133	0.0000	32	32
	β_3	-0.4472	-4.1586	0.9699	1	0	β_3	-0.4331	-4.2627	0.9660	1	0
Domestic Hybrid ($N = 77$)	α	-0.0736	-0.5640	0.6536	20	0	α	-0.0039	-0.0820	0.5289	34	3
	β_1	0.6103	12.9427	0.0018	77	76	β_1	0.6075	21.6223	0.0010	77	76
	γ	0.0129	0.1631	0.4542	43	3	γ	-0.0073	-0.1878	0.5344	35	0
	β_2	0.1063	2.6863	0.1465	66	50	β_2	0.1041	2.6682	0.1476	66	47
	β_3	0.0966	2.3280	0.2234	60	48	β_3	0.0948	2.4237	0.2298	60	48

Table 5

Cross-Sectional Distributions of Fund Alphas and Gammas – Summary Statistics

The table reports summary statistics of the cross-sectional distributions of estimated selection coefficients α and timing coefficients γ. The sample consists of 558 mutual funds and the estimation period is from January 1988 to March 1998 (123 monthly observations). For each of the four tests (HM Test, Equation 1; Adjusted Test, Equation 2; HM-FF3 Test, Equation 3; Adjusted-FF3 Test, Equation 4) we collect the distribution of estimated alphas and gammas. On the basis of each distribution we compute the mean, the standard deviation, minimum, maximum, the number of positive coefficients (denoted as $t > 0$), the number of positive coefficients that are at the same time statistically significant at the 5% level (denoted as $t > 0$ & $p < 0.05$), as well as the skewness and the degree of excess (kurtosis-3). Histograms for all eight coefficients are presented in Figure 1 (the CAPM-based tests) and in Figure 2 (tests based on the Fama-French 3-factor model).

	HM Test		Adjusted Test		HM-FF3 Test		Adjusted-FF3 Test	
	α	γ	α	γ	α	γ	α	γ
mean	0.0365	-0.0783	0.4519	-0.0745	-0.0656	0.0218	0.0609	-0.0136
standard deviation	0.3816	0.2039	0.6439	0.0928	0.3410	0.1611	0.4695	0.0615
minimum	-2.2238	-1.3086	-1.0292	-0.5103	-2.3009	-0.7453	-1.6774	-0.2912
maximum	1.2409	0.4206	3.4963	0.1189	1.2685	0.5386	2.3956	0.1636
$t > 0$	293	197	420	109	206	345	290	268
$t > 0$ & $p < 0.05$	32	16	60	2	14	31	19	12
skewness	-0.5698	-1.0100	1.0931	-1.2611	-0.7367	-0.9162	0.5709	-0.7932
degree of excess	3.8539	3.1185	1.7614	2.1363	7.9434	3.4393	2.7845	2.5086

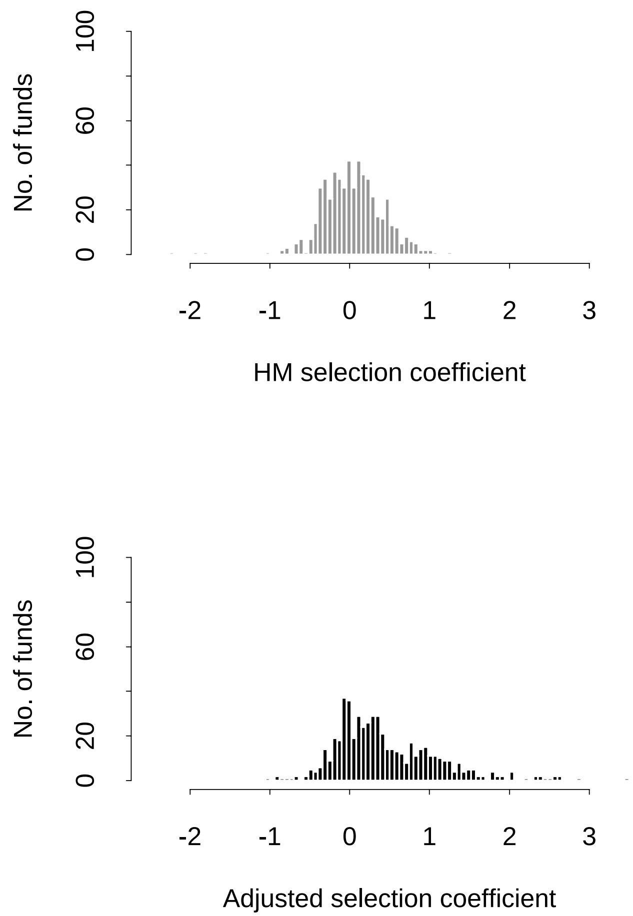

Figure 1: The figure displays cross-sectional distributions of the estimated selection coefficients α (expressed in percent per month) and timing coefficients γ under the two CAPM-based performance measures (HM Test, Equation 1; Adjusted Test, Equation 2). The sample of mutual funds consists of 558 mutual funds and the estimation period is from January 1988 to March 1998 (123 monthly observations). Summary statistics for all four coefficients featured in Figure 1 are presented in Table 5.

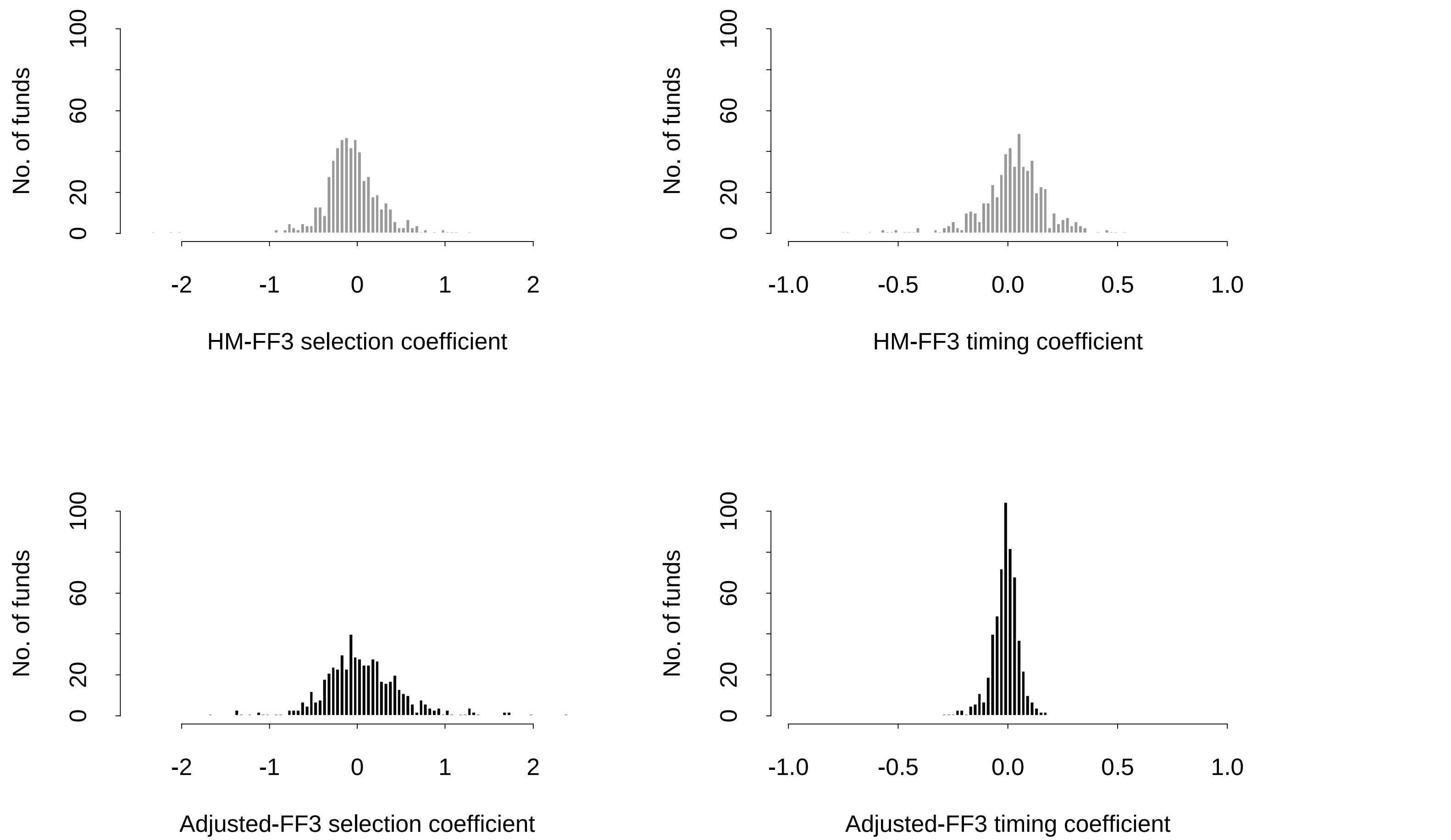

Figure 2: The figure displays cross-sectional distributions of the estimated selection coefficients α (expressed in percent per month) and timing coefficients γ under the two Fama-French-based performance measures (HM-FF3 Test 3, Equation 3; Adjusted-FF3 Test, Equation 4). The sample of mutual funds consists of 558 mutual funds and the estimation period is from January 1988 to March 1998 (123 monthly observations). Summary statistics for all four coefficients featured in Figure 2 are presented in Table 5.

Table 6
Performance of Timing Tests on Index Funds

The table reports the results of timing tests performed on the six index funds from our sample of 558 mutual funds in the period from January 1988 to March 1998 (123 monthly observations). The results are reported individually for each index fund. For each fund, the table reports the coefficient estimates and their respective statistical significance. Panel A displays the results of the classic HM test (Equation 1), Panel B displays the results of our adjusted test (Equation 2), Panel C displays the results of the HM-FF3 test (Equation 3), and Panel D displays the results of the adjusted-FF3 test (Equation 4). Our mutual fund sample and the data required to execute the four tests are discussed in Section IV.2.

	Panel A: HM Tests			**Panel C: HM-FF3 Tests**				
Fund Name	α	β	γ	α	β_1	γ	β_2	β_3
CoreFund Equity Index Y	−0.0979**	1.0021***	0.0124	−0.1229***	1.0115***	0.0214	−0.0049	0.0264***
SEI Index S&P 500 Index E	−0.0140	0.9970***	−0.0021	−0.0113	0.9957***	−0.0044	−0.0074***	0.0002
Stagecoach Equity Index A	−0.0422	0.9823***	−0.0322*	−0.0347	0.9790***	−0.0367	−0.0102	−0.0034
Vanguard Index 500	−0.0158**	1.0027***	0.0053*	−0.0178***	1.0034***	0.0055**	−0.0032**	0.0032**
Vanguard Index Extend Mkt	0.2515	0.8318***	−0.2019	−0.0578	0.9777***	0.0331	0.7190***	0.0210
Vanguard Index Sm Cap Stk	0.5638*	0.7351***	−0.4611**	0.0023	0.9922***	−0.0677	1.0989***	0.1187***

	Panel B: Adjusted Tests			**Panel D: Adjusted-FF3 Tests**				
Fund Name	α	β	γ	α	β_1	γ	β_2	β_3
CoreFund Equity Index Y	−0.1774***	0.9911***	0.0138**	−0.1811***	0.9969***	0.0126**	−0.0044	0.0242***
SEI Index S&P 500 Index E	0.0318	1.0009***	−0.0070	0.0409	1.0011***	−0.0083	−0.0084***	0.0005
Stagecoach Equity Index A	−0.1766**	0.9900***	0.0129	−0.1713**	0.9903***	0.0122	−0.0053	0.0013
Vanguard Index 500	−0.0104	1.0004***	0.0002	−0.0072	1.0010***	−0.0004	−0.0037***	0.0026*
Vanguard Index Extend Mkt	0.8083**	0.9655***	−0.1188**	0.0202	0.9648***	−0.0044	0.7157***	0.0170
Vanguard Index Sm Cap Stk	1.2193**	1.0029***	−0.1833**	−0.0306	1.0259***	−0.0090	1.1027***	0.1265***

* $p < 0.1$

** $p < 0.05$

*** $p < 0.01$

Table 7

Cross-Sectional Distributions of Stock Index Alphas and Gammas

The table reports summary statistics of the cross-sectional distributions of estimated selection coefficients α and timing coefficients γ. The sample consists of 55 stock indexes and the estimation period is from January 1988 to March 1998 (123 monthly observations). For each of the four tests (HM Test, Equation 1; Adjusted Test, Equation 2; HM-FF3 Test, Equation 3; Adjusted-FF3 Test, Equation 4) we collect the distribution of estimated alphas and gammas. On the basis of each distribution we compute the mean, the standard deviation, minimum, maximum, the number of coefficients that are different from zero at the 10% level of statistical significance (denoted as $p < 0.1$), as well as the skewness and the degree of excess (kurtosis-3) and report in the top part of the table. Moreover, we also provide a more detailed distribution of the coefficients according to their sign *and* statistical significance in the bottom part of the table.

	HM Test		Adjusted Test		HM-FF3 Test		Adjusted-FF3 Test	
	α	γ	α	γ	α	γ	α	γ
mean	0.1711	-0.1318	0.4441	-0.0666	-0.0601	0.0317	-0.0729	0.0088
standard deviation	0.3079	0.1938	0.5337	0.0857	0.1332	0.0767	0.1829	0.0251
minimum	-0.4757	-0.4976	-0.4196	-0.2415	-0.3586	-0.1227	-0.5853	-0.0424
maximum	0.8550	0.2550	1.7657	0.0561	0.2327	0.1975	0.3264	0.0679
$p < 0.1$	27	28	23	21	20	13	11	6
skewness	0.0049	0.0315	0.4838	-0.5942	-0.2385	0.3503	-0.3029	0.1522
degree of excess	-0.4153	-0.6976	-0.8104	-0.9000	-0.3115	-0.6382	0.3925	0.1515
$t < 0$ & $p < 0.01$	0	0	0	0	2	0	0	0
$t < 0$ & $0.01 < p < 0.05$	2	10	2	11	5	1	3	0
$t < 0$ & $0.05 < p < 0.1$	2	11	1	6	6	1	6	0
$t < 0$ & $p > 0.1$	10	18	10	22	24	22	32	18
$t > 0$ & $p > 0.1$	18	9	22	12	11	20	12	31
$t > 0$ & $0.05 < p < 0.1$	13	3	10	1	3	4	2	4
$t > 0$ & $0.01 < p < 0.05$	8	3	10	3	4	7	0	2
$t > 0$ & $p < 0.01$	2	1	0	0	0	0	0	0

Table 8
Performance of Asset Allocation Funds

The table reports the results of six timing tests performed on 23 Asset Allocation funds (from our sample of 558 mutual funds) in the period from January 1988 to March 1998 (123 monthly observations). The results are summarized by tests: the classic HM test (Equation 1), the adjusted test (Equation 2), the HM-FF3 test (Equation 3), the adjusted-FF3 test (Equation 4), the HM-5 test (Equation 5), and the adjusted-5 (Equation 6). The 23 funds analyzed in this table stated Asset Allocation as their Prospectus Objective. Twenty of them were classified as Domestic Hybrid funds. For each test, we report the average values of the respective estimated regression coefficients, the corresponding average t-statistics, the average p-value, the number of positive coefficients (denoted as $t > 0$) among the 23 funds, and the number of positive coefficients across all the funds that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). Our mutual fund sample and the data required to execute the six tests are discussed in Section IV.2.

Test (Equation No.)	Coefficient		t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
HM (1)	α	0.0930	0.5765	0.3911	13	5
	β	0.4599	8.3263	0.0063	23	21
	γ	-0.0430	-0.5514	0.5837	11	0
Adjusted (2)	α	0.2526	0.5773	0.3562	17	3
	β	0.4908	12.7745	0.0008	23	23
	γ	-0.0311	-0.4883	0.6196	7	0
HM-FF3 (3)	α	-0.0469	-0.1813	0.5418	9	0
	β_1	0.5181	9.3674	0.0061	23	22
	γ	0.0295	0.2460	0.4214	15	0
	β_2	0.1147	2.5325	0.1612	19	14
	β_3	0.0918	1.7500	0.2272	18	12
Adjusted-FF3 (4)	α	0.1013	0.2711	0.4257	14	2
	β_1	0.5112	14.4378	0.0034	23	22
	γ	-0.0151	-0.2747	0.5543	9	0
	β_2	0.1099	2.5100	0.1646	19	13
	β_3	0.0878	1.7342	0.2607	18	12
HM-5 (5)	α	-0.0624	-0.3535	0.5783	9	0
	β_1	0.4478	8.5313	0.0220	23	20
	γ	0.0430	0.4915	0.3637	16	2
	β_2	0.1634	3.3361	0.0792	21	17
	β_3	0.0514	1.2492	0.2634	17	9
	β_4	-0.0058	0.2789	0.4963	9	5
	β_5	0.2016	3.7167	0.1395	20	15
Adjusted-5 (6)	α	0.0959	0.2082	0.4337	14	2
	β_1	0.4363	10.3776	0.0264	22	22
	γ	-0.0137	-0.2113	0.5381	9	1
	β_2	0.1596	3.2951	0.0873	21	17
	β_3	0.0475	1.2251	0.2852	17	9
	β_4	-0.0120	0.2381	0.5211	8	5
	β_5	0.2014	3.7325	0.1397	20	15

Table 9
Timing Tests by Volatility of Implied Asset Weights

The table reports the results of the four timing tests (Equations 1 through 4) performed on some of the funds from our sample of 558 mutual funds in the period from January 1988 to March 1998 (123 monthly observations). The funds were included into this analysis if they belonged to either the top 10 percentile or the bottom 10 percentile of all the 558 funds with respect to the volatility of implied Sharpe weights (Sharpe, 1992). We computed the Sharpe implied weights for the S & P 500 and the 30-day Treasury Bill using a 12-month rolling window. We define the volatility of the resulting weight on the S & P 500 as the sum of absolute values of successive period weight changes. Results are summarized for the top 10 percentile, top 5 percentile, bottom 10 percentile, and bottom 5 percentile. For each of these groups of funds, we report the average values of the estimated regression coefficients, the corresponding average t-statistics, the average p-value, the number of positive coefficients (denoted as $t > 0$) among the funds from that group, and the number of positive coefficients across all the funds from that group that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). Panels A through D display results of the classic HM test (Equation 1), our adjusted test (Equation 2), the HM-FF3 test (Equation 3), and the adjusted-FF3 test (Equation 4), respectively. The mutual fund sample and the data required to execute the four tests are discussed in Section IV.2.

Panel A: HM Test by Volatility of Implied Asset Weights

Volatility Percentile (N)		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
	α	0.3147	0.7694	0.2795	45	11
Top 10% ($N = 56$)	β	0.6805	5.7194	0.0004	56	56
	γ	-0.2360	-0.8579	0.7464	8	1
	α	0.5027	1.2513	0.1724	26	8
Top 5% ($N = 28$)	β	0.6101	5.1420	0.0005	28	28
	γ	-0.3330	-1.2080	0.8328	2	0
	α	-0.1922	-0.9296	0.7411	10	0
Bottom 10% ($N = 56$)	β	1.1039	34.8282	0.0000	56	56
	γ	0.0555	0.4943	0.3678	36	6
	α	-0.1460	-0.8417	0.7204	7	0
Bottom 5% ($N = 28$)	β	1.1188	49.8813	0.0000	28	28
	γ	0.0292	0.2931	0.4156	17	1

Panel B: Adjusted Test by Volatility of Implied Asset Weights

Volatility Percentile (N)		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
	α	0.8649	1.1690	0.1777	51	18
Top 10% ($N = 56$)	β	0.8307	8.4554	0.0000	56	56
	γ	-0.1249	-1.1268	0.8259	6	0
	α	0.9827	1.3030	0.1476	26	11
Top 5% ($N = 28$)	β	0.8039	7.9107	0.0000	28	28
	γ	-0.1333	-1.1468	0.8325	3	0
	α	0.3808	0.5619	0.3240	41	6
Bottom 10% ($N = 56$)	β	1.1114	51.6898	0.0000	56	56
	γ	-0.0709	-0.9404	0.7646	8	1
	α	0.4370	0.5314	0.3223	21	3
Bottom 5% ($N = 28$)	β	1.1400	75.6331	0.0000	28	28
	γ	-0.0775	-0.9703	0.7661	5	1

Table 9 (continued)
Timing Tests by Volatility of Implied Asset Weights

Panel C: HM-FF3 Test by Volatility of Implied Asset Weights

Volatility Percentile (N)		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
Top 10% ($N = 56$)	α	-0.0374	-0.1085	0.5113	29	4
	β_1	0.8433	9.3516	0.0001	56	56
	γ	0.0176	0.1597	0.4550	32	7
	β_2	0.7325	11.1078	0.0305	54	52
	β_3	0.0575	0.9119	0.3628	36	23
Top 5% ($N = 28$)	α	0.0477	0.1841	0.4426	17	3
	β_1	0.8141	9.2520	0.0002	28	28
	γ	-0.0328	-0.0814	0.5124	15	2
	β_2	0.7739	12.6650	0.0574	26	25
	β_3	0.1417	1.7727	0.2384	22	14
Bottom 10% ($N = 56$)	α	-0.0867	-0.7134	0.6654	17	0
	β_1	1.0745	30.1048	0.0000	56	56
	γ	0.0638	0.6506	0.3302	42	5
	β_2	0.3076	4.5764	0.1178	50	46
	β_3	-0.2231	-2.6949	0.7945	12	3
Bottom 5% ($N = 28$)	α	-0.0369	-0.6244	0.6417	11	0
	β_1	1.0891	41.8649	0.0000	28	28
	γ	0.0403	0.4618	0.3706	18	1
	β_2	0.3337	4.8186	0.1482	24	22
	β_3	-0.2368	-2.8792	0.7865	6	2

Panel D: Adjusted-FF3 Test by Volatility of Implied Asset Weights

Volatility Percentile (N)		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
Top 10% ($N = 56$)	α	0.0480	0.0910	0.4572	32	1
	β_1	0.8390	14.5422	0.0000	56	56
	γ	-0.0086	-0.0494	0.4572	24	3
	β_2	0.7297	10.9532	0.0353	54	53
	β_3	0.0552	0.8834	0.3740	36	24
Top 5% ($N = 28$)	α	0.0847	0.1925	0.4192	18	1
	β_1	0.8336	14.9521	0.0000	28	28
	γ	-0.0119	-0.0952	0.5573	11	2
	β_2	0.7745	12.4525	0.0632	26	26
	β_3	0.1453	1.8330	0.2403	21	16
Bottom 10% ($N = 56$)	α	0.1358	0.1504	0.4445	32	3
	β_1	1.0539	50.5289	0.0000	56	56
	γ	-0.0188	-0.2654	0.5836	19	2
	β_2	0.2995	4.4272	0.1214	48	46
	β_3	-0.2313	-2.9708	0.8052	12	2
Bottom 5% ($N = 28$)	α	0.1658	0.0951	0.4536	15	2
	β_1	1.0797	74.6072	0.0000	28	28
	γ	-0.0207	-0.2590	0.5848	10	2
	β_2	0.3271	4.7797	0.1495	23	22
	β_3	-0.2422	-3.1320	0.7891	6	1

Table 10
Performance of "Glamour" Funds

The table reports the results of six timing tests performed on 43 "Glamour" funds (from our sample of 558 mutual funds) in the period from January 1988 to March 1998 (123 monthly observations). The results are summarized by tests: the classic HM test (Equation 1), the adjusted test (Equation 2), the HM-FF3 test (Equation 3), the adjusted-FF3 test (Equation 4), the HM-5 test (Equation 5), and the adjusted-5 (Equation 6). The 43 funds analyzed in this table were classified by Brown and Goetzmann (1997) as "Glamour" funds. For each test, we report the average values of the respective estimated regression coefficients, the corresponding average t-statistics, the average p-value, the number of positive coefficients (denoted as $t > 0$) among the 43 funds, and the number of positive coefficients across all the funds that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). Our mutual fund sample and the data required to execute the six tests are discussed in Section IV.2.

Test (Equation No.)		Coefficient	t-statistic	p-value	$t > 0$	$t > 0$ & $p < 0.05$
HM (1)	α	0.1122	0.2085	0.4314	27	0
	β	1.0361	8.9317	0.0000	43	43
	γ	-0.1465	-0.5203	0.6570	10	1
Adjusted (2)	α	1.2268	1.5067	0.1189	40	9
	β	1.1765	13.7778	0.0000	43	43
	γ	-0.1876	-1.7716	0.9108	2	0
HM-FF3 (3)	α	0.2308	0.5389	0.3571	30	2
	β_1	1.0213	10.6621	0.0000	43	43
	γ	-0.0582	-0.2523	0.5609	20	1
	β_2	0.8398	10.8755	0.0000	43	43
	β_3	-0.4440	-4.5015	0.9405	2	2
Adjusted-FF3 (4)	α	0.4686	0.6781	0.2971	35	4
	β_1	1.0658	17.8046	0.0000	43	43
	γ	-0.0456	-0.4580	0.6453	11	1
	β_2	0.8368	10.8955	0.0000	43	43
	β_3	-0.4384	-4.6639	0.9419	2	2
HM-5 (5)	α	0.3364	0.9018	0.2775	32	7
	β_1	1.0331	10.3951	0.0000	43	43
	γ	-0.1057	-0.5248	0.6359	14	0
	β_2	0.9301	10.0870	0.0000	43	43
	β_3	-0.4117	-3.8760	0.9307	2	2
	β_4	-0.2352	-2.0187	0.9031	2	0
	β_5	0.1011	1.0850	0.2351	37	10
Adjusted-5 (6)	α	0.5589	0.8755	0.2571	37	6
	β_1	1.0999	14.9871	0.0000	43	43
	γ	-0.0532	-0.5902	0.6747	11	1
	β_2	0.9279	10.0003	0.0000	43	43
	β_3	-0.4022	-3.9186	0.9297	2	2
	β_4	-0.2298	-2.0305	0.9016	1	0
	β_5	0.1016	1.0996	0.2327	37	11

Table 11
Effect of Survivorship

The table reports the results of six timing tests performed on 558 surviving funds (obtained from Morningstar) and 161 non-surviving funds (obtained from the CRSP Survivor Bias Free Mutual Fund Database) for horizons of k years ($k = 2, \ldots, 7$) with the beginning of each period in January of 1988. Within each horizon, the results are summarized by tests: the classic HM test (Equation 1), the adjusted test (Equation 2), the HM-FF3 test (Equation 3), the adjusted-FF3 test (Equation 4), the HM-5 test (Equation 5), and the adjusted-5 (Equation 6). For each test, we report the average values of the selection and timing coefficients and the number of positive coefficients across all the funds that are at the same time statistically significant (denoted as $t > 0$ & $p < 0.05$). For illustrative purposes, we also provide t-statistics of simple comparisons of mean values of alphas and gammas (denoted as $t(\Delta\alpha)$ and $t(\Delta\gamma)$, respectively). Note that the t-tests should be interpreted merely as an illustrative tool because they do not take into account correlations among the coefficient estimates. Our mutual fund sample and the data required to execute the six tests are discussed in Section IV.2.

Horizon (k years) **Test** (Equation No.)	α	$t > 0$ & $p < 0.05$	γ	$t > 0$ & $p < 0.05$	α	$t > 0$ & $p < 0.05$	γ	$t > 0$ & $p < 0.05$	$t(\Delta\alpha)$	$t(\Delta\gamma)$
$k = 2$	**Surviving**(N=558)				**Non-surviving**(N=161)					
HM (1)	-0.2844	9	0.2031	26	-0.3689	3	0.0832	9	1.6100	3.8372
Adjusted (2)	0.6118	126	-0.0838	67	0.2214	36	-0.0629	15	2.2493	-0.9298
HM-FF3 (3)	-0.0804	14	0.0757	44	-0.2499	2	-0.0059	12	3.1072	2.6926
Adjusted-FF3 (4)	0.1374	59	-0.0172	74	0.0407	20	-0.0409	15	0.7060	1.3620
HM-5 (5)	-0.1021	8	0.0710	44	-0.2692	0	-0.0050	12	3.1580	2.5319
Adjusted-5 (6)	0.2050	47	-0.0312	34	0.0859	13	-0.0509	10	0.9197	1.1481
$k = 3$	**Surviving**(N=558)				**Non-surviving**(N=130)					
HM (1)	-0.0603	8	-0.0891	15	-0.2065	1	-0.0994	4	3.1882	0.3925
Adjusted (2)	0.8379	102	-0.1236	20	0.6328	21	-0.1186	3	1.3342	-0.2684
HM-FF3 (3)	-0.0779	12	0.0579	49	-0.2394	1	0.0345	13	3.5337	1.0343
Adjusted-FF3 (4)	-0.1164	23	0.0154	83	-0.0343	9	-0.0185	17	-0.7151	2.4416
HM-5 (5)	-0.0766	19	0.0406	57	-0.2365	1	0.0204	15	3.4800	0.8439
Adjusted-5 (6)	-0.1254	21	0.0130	69	-0.0655	8	-0.0172	14	-0.5378	2.2195
$k = 4$	**Surviving**(N=558)				**Non-surviving**(N=103)					
HM (1)	0.1122	25	-0.0958	19	-0.1020	1	-0.0630	3	4.1225	-1.1506
Adjusted (2)	0.8139	91	-0.1037	11	0.4465	17	-0.0787	3	2.3682	-1.3798
HM-FF3 (3)	-0.0351	22	0.0327	41	-0.2378	1	0.0554	9	4.2930	-0.9866
Adjusted-FF3 (4)	-0.0579	19	0.0089	54	-0.2203	7	0.0083	14	1.4410	0.0432
HM-5 (5)	-0.0089	29	0.0206	48	-0.2085	1	0.0429	10	4.0561	-0.9465
Adjusted-5 (6)	-0.0194	23	0.0052	59	-0.2076	5	0.0081	13	1.6833	-0.2167

Table 11 (continued)
Effect of Survivorship

Horizon (k years) **Test** (Equation No.)	α	$t > 0$ & $p < 0.05$	γ	$t > 0$ & $p < 0.05$	α	$t > 0$ & $p < 0.05$	γ	$t > 0$ & $p < 0.05$	$t(\Delta\alpha)$	$t(\Delta\gamma)$
$k = 5$	**Surviving**(N=558)				**Non-surviving**(N=161)					
HM (1	0.1833	34	-0.1247	17	-0.0987	1	-0.0802	1	4.9226	-1.3733
Adjusted (2)	0.8573	99	-0.1119	5	0.5311	9	-0.0976	2	2.3464	-0.8111
HM-FF3 (3)	0.0301	22	0.0148	24	-0.2174	0	0.0377	2	4.9704	-0.9868
Adjusted-FF3 (4)	0.1347	40	-0.0108	30	-0.1472	3	-0.0018	7	2.9102	-0.7497
HM-5 (5)	0.0529	38	0.0051	31	-0.1891	0	0.0262	2	4.5951	-0.8864
Adjusted-5 (6)	0.1818	49	-0.0159	35	-0.1089	4	-0.0053	5	2.9209	-0.8732
$k = 6$	**Surviving**(N=558)				**Non-surviving**(N=51)					
HM (1)	0.2324	79	-0.1397	16	-0.1080	1	-0.0490	2	5.1158	-2.3493
Adjusted (2)	0.8435	140	-0.1116	2	0.4562	6	-0.0879	0	2.7521	-1.2643
HM-FF3 (3)	0.0593	29	0.0103	25	-0.1993	0	0.0555	1	4.6221	-1.6819
Adjusted-FF3 (4)	0.2014	38	-0.0178	27	-0.0731	1	-0.0067	2	2.7189	-0.8535
HM-5 (5)	0.0803	47	-0.0003	27	-0.1803	0	0.0462	2	4.3130	-1.6770
Adjusted-5 (6)	0.2307	52	-0.0209	34	-0.0443	0	-0.0096	3	2.5571	-0.8390
$k = 7$	**Surviving**(N=558)				**Non-surviving**(N=36)					
HM (1)	0.1487	48	-0.1114	18	-0.2230	1	-0.0227	2	5.1331	-2.0774
Adjusted (2)	0.6512	107	-0.0948	1	0.2253	2	-0.0696	0	2.9165	-1.2481
HM-FF3 (3)	-0.0067	17	0.0180	26	-0.2807	0	0.0513	0	4.5093	-1.0973
Adjusted-FF3 (4)	0.0658	24	-0.0068	31	-0.2699	0	0.0089	2	3.3649	-1.1387
HM-5 (5)	0.0169	32	0.0076	28	-0.2626	0	0.0434	0	4.3294	-1.1599
Adjusted-5 (6)	0.0902	28	-0.0090	38	-0.2494	0	0.0070	3	3.2846	-1.1463

APPENDIX

Fifty-five Stock Indexes

The table lists the 55 stock indexes used in the analyses reported in Table 7. The stock indexes were available for the period from January 1988 to March 1998 from Ibbotson Associates.

S&P/BARRA 500 Growth	FT/S&P U.S. Large Cap
S&P/BARRA 500 Value	FT/S&P U.S. Med-Small Cap
Wilshire Large Growth	U.S. Small Stk
Wilshire Large Value	BGI Extended Equity
Wilshire Small Growth	BGI Interm Cap Growth
Wilshire Small Value	BGI Interm Cap
Wilshire MidCap Growth	BGI Interm Cap Value
Wilshire MidCap Value	BGI Medium Cap Growth
Russell 3000 Growth	BGI Medium Cap
Russell 3000 Value	BGI Medium Cap Value
Russell 1000 Growth	BGI Small Cap Growth
Russell 1000 Value	BGI Small Cap
Russell 2000 Growth	BGI Small Cap Value
Russell 2000 Value	BGI Micro Cap
S&P500	S&P MidCap 400
Russell Top 200 Growth	Wilshire Top 750
Russell Top 200 Value	Wilshire Next 1750
Russell MidCap Growth	Wilshire 4500
Russell MidCap Value	Wilshire 5000
Russell 2500 Growth	Russell 1000
Russell 2500 Value	Russell 2000
IIA U.S. Growth	Russell 3000
IIA U.S. Large Cap Growth	S&P SmallCap 600
IIA U.S. Large Cap Value	Russell 2500
IIA U.S. Small Cap Growth	Russell MidCap
IIA U.S. Small Cap Value	Russell Top 200
IIA U.S. Value	
IIA U.S. Large Cap	
IIA U.S. Small Cap	