



ELSEVIER

Journal of Financial Economics 62 (2001) 377–411

JOURNAL OF  
Financial  
ECONOMICS

www.elsevier.com/locate/econbase

# The performance of professional market timers: daily evidence from executed strategies<sup>☆</sup>

Don M. Chance<sup>a</sup>, Michael L. Hemler<sup>b,\*</sup>

<sup>a</sup> *Department of Finance, Virginia Tech, Blacksburg, VA 24061, USA*

<sup>b</sup> *Department of Finance and Business Economics, University of Notre Dame, Notre Dame, IN 46556, USA*

Received 18 May 1998; received in revised form 6 February 2001

---

## Abstract

We examine the performance of 30 professional market timers during 1986–1994. Prior studies have analyzed implicit recommendations from mutual fund returns or explicit recommendations from newsletters. We analyze explicit recommendations executed in customer accounts. Using four tests, three benchmark portfolios, and daily data, we find significant unconditional and conditional ability that is robust with respect to transaction costs and survivorship bias. Relative ability persists and varies with the frequency of recommendation changes. When recommendations of successful timers are observed monthly instead of daily, significant ability generally disappears. Hence, the

---

<sup>☆</sup>The authors gratefully acknowledge financial support from Virginia Tech and the University of Notre Dame. We thank Champion Securities LP, MoniResearch Corporation, and Rob Bliss for graciously providing data used in this research. We also thank Bill Schwert (the editor), an anonymous referee, John Affleck-Graves, Rob Bliss, Bob Champion, Bent Jesper Christensen, Wayne Ferson, Ken French, Will Goetzmann, Bruce Grundy, Scott Harrington, Campbell Harvey, Peter Jorgensen, Greg Kadlec, Fred Lindahl, Rick Mendenhall, Andrew Metrick, Wayne Mikkelson, Cathy Niden, Greg Niehaus, Gary Porter, Paul Schultz, Richard Sheehan, Steve Shellans, David Upton, as well as seminar participants at the University of Aarhus, the University of Notre Dame, the University of South Carolina, Virginia Tech, the 1998 Financial Management Association meeting, the 1999 Eastern Finance Association meeting, and the 1999 Western Finance Association meeting for helpful comments.

\*Corresponding author. Tel.: +1-219-631-6766; fax: +1-219-631-5255.

*E-mail address:* mhemler@nd.edu (M.L. Hemler).

0304-405X/01/\$ - see front matter © 2001 Elsevier Science S.A. All rights reserved.  
PII: S 0 3 0 4 - 4 0 5 X ( 0 1 ) 0 0 0 8 1 - 2

frequency with which recommendations are observed can change inferences regarding ability. © 2001 Elsevier Science S.A. All rights reserved.

*JEL classification:* G11; G14; G23

*Keywords:* Market timing; Performance evaluation; Portfolio management; Asset allocation

---

## 1. Introduction

This paper investigates the performance of 30 professional market timers during 1986–1994. A notable feature of this analysis is its relatively unique data. Previous studies typically have used implicit recommendations estimated from mutual fund returns or explicit recommendations obtained from investment newsletters. This study uses explicit recommendations executed in customer accounts. These timers have limited power of attorney, indicating that their clients trust their ability. They disclose their recommendations voluntarily, suggesting self-confidence in their performance. If any timers possess significant ability, the timers studied here are likely candidates.

We use four tests to analyze performance on a daily basis: mean-variance tests, unconditional and conditional Cumby and Modest (1987) regression tests, and Graham and Harvey (1996) weight change tests. Each test uses benchmark portfolios that correspond to Nasdaq, NYSE/Amex/Nasdaq, and the Standard & Poor's (S&P) 500.

We find evidence of significant ability across all tests and portfolios. This evidence is robust with respect to transaction costs and survivorship bias. We also find that relative performance persists and varies with the frequency at which a timer changes recommendations. For comparison purposes, we examine a consensus timer, whose recommendations reflect forecasts based on all 30 timers, and a statistical timer, whose recommendations reflect forecasts based on standard macroeconomic variables. The consensus timer demonstrates both unconditional and conditional ability. Because the statistical timer forecasts based on conditioning variables, we expect it to show only unconditional ability, and that is the case. Finally, we obtain an interesting new result involving the frequency with which recommendations are observed. When we reconsider our most successful timers, but update their recommendations monthly instead of daily as done originally, we find that their significant ability vanishes. Thus, inferences regarding timing ability can vary dramatically depending on how frequently recommendations are observed.

The paper proceeds as follows. Section 2 discusses prior research and methodological issues. Section 3 describes the data, and Section 4 presents initial results from various market timing tests. Section 5 examines timing ability in more detail, exploring topics such as consensus and statistical

forecasting, survivorship bias, persistence, transaction costs and management fees, the frequency with which a timer changes recommendations, and the frequency with which a researcher observes recommendations. Section 6 offers concluding remarks.

## 2. Market timing: major issues and prior research

What is successful market timing? Consider a one-period world and assume that there are only two asset classes: stocks, as represented by a broad-based index, and cash, as represented by a money market fund. Assume that a portfolio manager invests 100% in stocks and 0% in cash or vice versa. Successful timing depends on the manager's ability to predict whether stocks will earn a higher return than cash. Given that prediction, a manager invests 100% in the asset class with the higher expected return.

To quantify timing ability, let the equity risk premium  $R_S - R_B$  denote the return on stocks less the return on cash. Let the indicator variable  $I$  equal one if  $R_S - R_B$  is positive and zero otherwise. Let the portfolio weight  $W$  denote the fraction of capital that a manager invests in stocks. Assume for simplicity that  $W$  equals either one or zero. If a manager has perfect timing ability, the correlation between  $I$  and  $W$  is one. If a manager has no timing ability and chooses  $W$  randomly without regard to  $R_S - R_B$ , the correlation is zero. We therefore equate successful timing ability with positive correlation in which the equity risk premium  $R_S - R_B$  replaces  $I$  and the portfolio weight  $W$  satisfies  $0 \leq W \leq 1$ . The intuition is that the equity risk premium varies positively with stock market exposure for market timers having significant ability.

This link between correlation and market timing ability is well-known. Merton (1981), Henriksson and Merton (1981), and Cumby and Modest (1987) all have exploited it. Merton (1981) provides a theoretical framework and analysis. Henriksson and Merton (1981) propose empirical tests, including a contingency table test. Cumby and Modest (1987) extend this contingency table test to a regression test and use both tests to study foreign exchange advisory services. More recently, Grinblatt and Titman (1989, 1994), Graham and Harvey (1994, 1996, 1997); Ferson and Schadt (1996), and Ferson and Warther (1996) utilize this link to varying degrees.<sup>1</sup>

Previous empirical tests of market timing have often utilized implicit recommendations inferred from estimated shifts in the risk of mutual funds.

---

<sup>1</sup>In addition, these recent papers examine conditional market timing, which differs significantly from the unconditional version just proposed. We defer discussion of conditional tests until Section 4.2, where we consider conditional and unconditional versions of the Cumby-Modest (1987) regression tests.

These tests generally conclude that mutual fund managers exhibit little timing ability.<sup>2</sup> Such tests, however, do not examine the exact timing decisions taken by fund managers. They must infer those decisions based on changes in fund risk, which introduces potential measurement error. For instance, Alexander et al. (1982) show that risk can change significantly even when no timing is undertaken. Thus, changes in risk need not mirror timing decisions.

An alternative and more direct approach for obtaining timer recommendations is to use investment newsletters. For example, Graham and Harvey (1994, 1996, 1997), examine newsletters tracked by *Hulbert Financial Digest*, an independent publication that rates newsletter performance. They conclude that newsletters generally perform poorly. Only 22.8% of the newsletters have average returns higher than a passive portfolio of equity and cash with the same volatility. Poor performance persists far more than good performance, but newsletters that are on a hot streak may provide valuable market timing information.<sup>3</sup>

Although newsletter recommendations are more explicit than recommendations extracted from mutual fund returns, they still possess disadvantages. Timers who write newsletters need not actually invest capital according to their recommendations. Timing itself could be a problem. A week could easily pass from the time a newsletter is written until the time a newsletter is received. Furthermore, a timer could change a recommendation after sending it. Clients who pay a premium could receive “hot-line” phone calls notifying them of the change; other clients would not.

In contrast, our study examines data from customer statements to determine the positions taken by market timers. One other paper studies known executed recommendations. Wagner et al. (1992) use data from MoniResearch Corporation (which also provides our data) to conclude that market timers exhibited superior performance from October 1985 to September 1990. This conclusion, however, relies on security market line estimates of alphas. Without adjusting for changing risk, one cannot separate security selection from market timing.<sup>4</sup> Moreover, as noted by Brocato and Chandy (1994), the conclusion could reflect survivorship bias. Our analysis differs from Wagner et al. (1992) by

<sup>2</sup> Representative studies include Treynor and Mazuy (1966), Kon and Jen (1978, 1979), Kon (1983), Chang and Lewellen (1984), Lehmann and Modest (1987), Lee and Rahman (1990), Grinblatt and Titman (1994), and Ferson and Schadt (1996).

<sup>3</sup> Other recent papers that examine newsletter performance include Graham (1999), Jaffe and Mahoney (1999), and Metrick (1999).

<sup>4</sup> Consider two timers. One invests 100% in stocks and the other invests 100% in cash. Suppose stocks earn 10% and cash earns 7% over the period. These timers each have an ex post alpha of zero, but their timing performances clearly differ. The problem lies in using traditional security line measures. Without adjusting for changes in risk, timing cannot be separated from selectivity. See Dybvig and Ross (1985), Admati et al. (1986), Grinblatt and Titman (1989, 1994), and Elton and Gruber (1991), for example.

using more appropriate tests and examining several additional topics such as survivorship bias. More generally, it differs from all prior research by using daily data to study known executed timing recommendations.

### **3. The data**

Our timer data consist of signals for 30 professional market timers, all Registered Investment Advisers, during 1986–1994. By the term “signal,” we mean an explicit timing recommendation. We distinguish between explicit and implicit recommendations. Upon giving a signal, a timer maintains that recommendation implicitly until revising it via a new signal.

As noted earlier, these data come from MoniResearch Corporation, which has been publishing a newsletter monitoring the performance of professional market timers since 1986. Although some timers might have provided recommendations prior to 1986, our data go back only to 1986, the newsletter’s first year of publication. MoniResearch also monitors asset allocators who invest in equity, cash, and other asset classes, but we restrict our attention to market timers; i.e., those who use only equity and cash. By virtue of a limited power of attorney, these timers usually invest part of their clients’ capital in a no-load mutual fund that tracks a stock market index and the rest in a money market fund. MoniResearch typically obtains client statements from mutual fund accounts for the monitored timers. It then extracts timing signals from these statements and verifies the dates on which the signals were implemented. No hypothetical or backfitted signals are permitted.

Our sample consists of some timers who ask MoniResearch to monitor their performance and others who manage sufficient capital that MoniResearch contacts them and requests their participation. MoniResearch refused to track a timer only if it thought the timer was dishonest or if the timer could not provide customer statements necessary for verifying transactions. MoniResearch did not charge the timers a fee during the period examined in this study, though in more recent years, a fee has been assessed. In regard to survivorship bias, there were ten timers who dropped out of the database during 1991–1994. No timers, however, left our sample and then reentered later under a different alias. Thus, there is no bias similar to the “re-emerging market” bias examined by Goetzmann and Jorion (1999).<sup>5</sup>

According to Steve Shellans, president of MoniResearch, these timers typically use small capitalization or growth funds instead of funds that track

---

<sup>5</sup> MoniResearch has confirmed that some overlap exists between timers examined here and in Graham and Harvey (1996). Three timers appear in both studies. But even when both studies analyze the same timer over the same period, corresponding recommendations need not coincide. A timer could give one set of recommendations in a newsletter but use another set of recommendations in a managed account.

the S&P 500. Because their performance is often judged relative to the S&P 500, we conjecture that these timers might believe they can game the evaluation process by utilizing funds that have higher levels of risk and expected return than the S&P 500. Alternatively, timers might be trying to exploit the relatively higher return autocorrelation present in small capitalization stocks, which makes small capitalization stock returns relatively more predictable than large capitalization stock returns.

A timer's preference for a more volatile asset class over a less volatile asset class is also consistent with a well-known theory of market timing. Merton (1981) shows that successful market timing provides a protective put, wherein the portfolio is fully invested in stocks when the market is going up and fully invested in cash when the market is going down. Given that a put option is more valuable the greater the volatility, a put on a more volatile asset class, such as small capitalization or growth stocks, is more valuable than a put on a less volatile asset class, such as S&P 500 stocks. Given that successful market timing provides a put option, it follows that a market timer who believes that he or she will be successful will prefer a more volatile asset class when choosing which asset class to use for the market. We might, therefore, expect that a market timer who believes he has ability would be confident that this ability would be upheld in an ex post analysis.

Table 1 describes the recommendations given by the timers in our sample. Each signal, expressed as the percentage of capital invested in stocks, ranges from 0% to 100%. Fifteen timers always recommend exactly one asset class at any time. There is considerable variation in the frequency with which timers change recommendations. The average time between signals ranges from 6.2 to 143.5 days. This mean exceeds the corresponding median for all but two timers, reflecting skewness typically present in the underlying distribution.

In addition to our timer data, we use daily stock and bond returns from the Center for Research in Security Prices (CRSP). The stock returns, obtained from the CRSP Indices/Excess Returns File, are returns, including dividends, for three value-weighted portfolios: Nasdaq, NYSE/Amex/Nasdaq, and the S&P 500. We use three portfolios to check whether inferences depend on the proxy used for "the market."<sup>6</sup> The S&P 500 portfolio is a popular benchmark

---

<sup>6</sup>We do not measure performance based on returns from the actual instruments used by these timers. There are two major reasons for this. First, we do not have the necessary information. MoniResearch does not record the actual instruments used by the timers. Second, if one changes the instrument depending on the timer, performance would depend not only on timing (the decision of whether to invest in stocks or cash), but also on selectivity (the decision of which fund to use). Different mutual funds, for instance, could reflect different investment styles, which could have an impact on timer performance. Performance measurement could also be corrupted by the timing and selectivity ability of the managers who run the mutual funds. Consequently, measuring performance relative to a benchmark common to all timers is the only way to accurately measure timer ability.

for measuring portfolio performance and determining managerial compensation. The NYSE/Amex/Nasdaq portfolio is the broadest portfolio, containing all stocks in the other two portfolios. The Nasdaq portfolio contains the highest percentage of small capitalization or growth stocks. Nearly all our timers use mutual funds based on such stocks.<sup>7</sup> The bond returns are risk-free rates for one-month Treasury bills obtained from the CRSP Government Bond File.<sup>8</sup>

#### 4. Tests of market timing ability

##### 4.1. Mean-standard deviation tests

Consider two returns for any given timer and sample period. The first is the realized return. The second is a matched return that corresponds to a strategy having the same standard deviation of return as the first return, but with constant portfolio weights relative to stocks and cash. To test whether a timer has ability, one can check whether the average realized return exceeds the average matched return. Graham and Harvey (1994, 1997), for example, use this test to study investment newsletter performance.

Table 2 gives estimated means  $\mu(R)$  and  $\mu(M)$  of realized and matched returns, in addition to their common standard deviation  $\sigma$ . It also provides the performance measure  $[\mu(R) - \mu(M)]/\sigma$ , which is denoted by *Ratio*.<sup>9</sup> The estimates in Table 2, however, do not allow for testing of whether  $\mu(R) > \mu(M)$  at any significance level. The standard *t*-test is inappropriate for at least two reasons: (1) realized and matched returns are not independent, and (2) realized returns, which are mixtures of stock and bond returns, are not normally distributed. Nonetheless, the estimate of  $\mu(R)$  typically exceeds the estimate of  $\mu(M)$ . This occurs 26 times for Nasdaq, 22 times for NYSE/Amex/Nasdaq, and 18 times for the S&P 500. Timers 12, 14, 23, 25, 26, and 27 give the best performance, and they change recommendations more frequently than any other timers in our sample.

<sup>7</sup>Kester (1990) finds that small capitalization stocks offer more profitable opportunities for market timing than large capitalization stocks during the period 1934–1988. Similar to Kester, we find that ability tends to appear strongest relative to Nasdaq and weakest relative to the S&P 500.

<sup>8</sup>Although these rates are provided in the Quarterly and Annual CRSP Government Bond Files, they are not provided in the Daily Files. We thank Rob Bliss, who generously calculated and provided these rates.

<sup>9</sup>One can interpret this measure as the difference between Sharpe ratios for two portfolios, assuming the Sharpe ratio is being used as an ex post measure wherein the numerator is the average daily return minus the average daily risk-free rate. Alternatively, it equals the performance measure GH1, i.e.,  $[\mu(R) - \mu(M)]$ , proposed by Graham and Harvey (1997), divided by the corresponding standard deviation.

Table 1

Characteristics of recommendations given by 30 professional market timers during 1986–1994. The term “signal” refers to an explicit market timing recommendation by a given timer; i.e., the first recommendation or a subsequent revision. Each timer gives daily recommendations, explicitly or implicitly, from the time of his first signal until the end of 1994. The data are from MoniResearch Corporation, which publishes a newsletter that evaluates professional market timer performance.

Timer <sup>a</sup>	Date of first signal	Date of last signal	Total number of signals	Number of signals at 0%	Number of signals between 0% and 100%	Number of signals at 100%	Median number of calendar days between signals	Average number of calendar days between signals	Standard deviation of number of calendar days between signals <sup>b</sup>
1	1/2/86	11/30/93	22	11	0	11	77.0	137.6	141.0
2	1/2/86	12/13/94	44	22	0	22	64.0	76.0	48.2
3	1/2/86	3/29/94	69	12	41	16	18.5	44.2	60.3
4	1/2/86	12/30/93	26	13	1	12	44.0	116.8	133.4
5	3/3/86	10/20/94	42	21	1	20	64.0	76.9	62.8
6	3/3/86	11/25/94	89	5	72	12	29.0	36.2	32.7
7	5/1/86	12/21/94	23	11	0	12	106.5	143.5	122.2
8	5/1/86	12/22/94	66	10	50	6	27.0	48.6	57.8
9	5/1/86	3/30/94	26	13	0	13	71.0	115.6	222.3
10	2/2/87	12/19/94	26	13	1	12	70.0	115.1	107.0
11	4/1/87	3/30/94	62	30	4	28	29.0	41.9	40.0
12	4/4/88	10/20/94	99	49	3	47	7.0	24.4	64.3
13	4/4/88	10/14/93	24	12	3	9	39.0	87.8	105.1
14	7/3/89	12/30/94	303	152	0	151	5.0	6.6	6.4
15	11/1/89	12/05/94	34	17	0	17	41.0	56.4	45.5
16	4/2/90	12/09/94	15	5	7	3	87.5	122.3	122.7
17	12/3/90	12/29/94	84	29	42	13	13.0	17.9	15.5
18	1/2/91	11/21/94	41	20	0	21	17.0	35.5	46.9
19	2/1/91	3/17/94	9	4	0	5	93.5	142.5	112.3
20	3/1/91	10/24/94	52	26	1	25	21.0	26.1	23.0
21	8/1/91	12/22/94	24	9	5	10	49.0	53.9	36.9
22	8/1/91	12/21/94	17	8	1	8	68.0	77.4	54.5

23	8/1/91	12/13/94	71	35	0	36	6.5	17.6	26.7
24	4/1/92	12/15/94	12	6	0	6	74.0	89.8	75.1
25	8/3/92	12/21/94	77	30	16	31	8.0	11.4	11.2
26	8/3/92	12/28/94	106	53	0	53	6.0	8.4	6.9
27	12/1/92	12/28/94	123	62	0	61	5.0	6.2	5.6
28	1/4/93	12/14/94	13	6	0	7	60.0	59.1	55.3
29	8/2/93	11/21/94	34	17	0	17	14.0	14.4	10.0
30	12/1/93	2/18/94	2	1	0	1	79.0	79.0	—

<sup>a</sup>Timer 15 gave two sets of recommendations depending on whether the market corresponds to the Standard & Poor's (S&P) 500 or Nasdaq. To have one set of recommendations per timer, we use the recommendations corresponding to the S&P 500. The rationale for choosing the S&P 500 over Nasdaq is that the S&P 500 is the more popular and familiar benchmark. In addition, timers 3 and 6 used only stocks and cash during this period, but they were not restricted to using only stocks and cash. They could invest in other asset classes.

<sup>b</sup>Because the number of calendar days between signals often has a highly skewed distribution, these standard deviations are for descriptive purposes only. No standard deviation appears for timer 30 because the usual estimator is undefined.

Table 2

Sample means and standard deviations of realized and matched daily returns relative to the Nasdaq, NYSE/Amex/Nasdaq, and Standard & Poor's (S&P) 500 stock portfolios. The realized returns correspond to the returns on the portfolio determined by recommendations for a given timer. The matched returns have the same standard deviation as the realized returns but, unlike the realized returns, have constant portfolio weights relative to stocks and cash. The proxy for "stocks" is either the Nasdaq, NYSE/Amex/Nasdaq, or S&P 500 stock portfolio. The proxy for "cash" is the return on a one-month Treasury bill. All means and standard deviations are realized values as opposed to expected values, and all values are nonannualized percentages. The means of the realized and matched returns are represented by  $\mu(R)$  and  $\mu(M)$ , respectively. The standard deviation of the realized (alternatively, matched) return is represented by  $\sigma$ . The performance measure  $[\mu(R) - \mu(M)]/\sigma$  is represented by *Ratio*, which can be interpreted as the difference between Sharpe ratios for two portfolios. Alternatively, it equals the performance measure GH1 proposed by Graham and Harvey (1997); i.e.,  $[\mu(R) - \mu(M)]$ , divided by the corresponding standard deviation. The number of observations used to calculate the relevant test statistics is represented by  $N$ . The sample period for each timer begins on the day of his first signal and ends on December 30, 1994.

Timer	$N$	Nasdaq				NYSE/Amex/Nasdaq				S&P 500			
		$\mu(R)$	$\mu(M)$	$\sigma$	<i>Ratio</i>	$\mu(R)$	$\mu(M)$	$\sigma$	<i>Ratio</i>	$\mu(R)$	$\mu(M)$	$\sigma$	<i>Ratio</i>
1	2249	0.064	0.035	0.607	0.048	0.052	0.036	0.609	0.027	0.051	0.039	0.698	0.017
2	2249	0.058	0.033	0.550	0.046	0.039	0.035	0.566	0.008	0.037	0.038	0.651	-0.002
3	2249	0.060	0.033	0.566	0.048	0.052	0.035	0.577	0.030	0.052	0.038	0.660	0.022
4	2249	0.023	0.033	0.553	-0.018	0.026	0.036	0.617	-0.018	0.028	0.039	0.695	-0.016
5	2209	0.050	0.026	0.415	0.058	0.041	0.029	0.460	0.026	0.040	0.032	0.535	0.015
6	2209	0.052	0.029	0.522	0.043	0.046	0.031	0.510	0.029	0.046	0.034	0.581	0.021
7	2167	0.057	0.028	0.525	0.056	0.047	0.031	0.538	0.029	0.045	0.034	0.612	0.019
8	2167	0.032	0.026	0.441	0.013	0.026	0.029	0.457	-0.005	0.026	0.031	0.525	-0.011
9	2167	0.048	0.030	0.615	0.029	0.032	0.032	0.580	-0.001	0.028	0.035	0.657	-0.012
10	1979	0.020	0.032	0.615	-0.019	0.020	0.031	0.618	-0.018	0.021	0.034	0.707	-0.018
11	1938	0.060	0.026	0.503	0.067	0.048	0.027	0.498	0.044	0.047	0.029	0.570	0.031
12	1687	0.104	0.039	0.590	0.111	0.087	0.037	0.504	0.100	0.088	0.039	0.569	0.087

13	1687	0.047	0.038	0.574	0.015	0.038	0.036	0.483	0.004	0.037	0.038	0.540	-0.002
14	1374	0.222	0.031	0.496	0.386	0.190	0.027	0.416	0.392	0.194	0.029	0.475	0.346
15	1290	0.064	0.034	0.607	0.050	0.050	0.028	0.473	0.048	0.049	0.028	0.519	0.039
16	1187	0.052	0.037	0.609	0.024	0.038	0.029	0.472	0.021	0.038	0.029	0.513	0.018
17	1019	0.046	0.041	0.385	0.014	0.043	0.029	0.293	0.050	0.045	0.028	0.317	0.054
18	999	0.028	0.037	0.353	-0.025	0.015	0.026	0.265	-0.041	0.011	0.025	0.283	-0.048
19	978	0.055	0.055	0.682	0.000	0.036	0.039	0.531	-0.005	0.033	0.037	0.576	-0.006
20	959	0.068	0.034	0.421	0.082	0.044	0.024	0.312	0.066	0.039	0.023	0.347	0.046
21	853	0.056	0.038	0.584	0.030	0.027	0.026	0.415	0.004	0.022	0.024	0.447	-0.006
22	853	0.060	0.038	0.571	0.039	0.026	0.025	0.404	0.002	0.018	0.024	0.438	-0.015
23	853	0.116	0.042	0.666	0.111	0.078	0.028	0.484	0.103	0.071	0.027	0.529	0.082
24	687	0.022	0.027	0.613	-0.009	0.017	0.024	0.449	-0.017	0.016	0.026	0.487	-0.019
25	603	0.064	0.028	0.416	0.085	0.042	0.018	0.304	0.078	0.036	0.017	0.328	0.058
26	603	0.119	0.027	0.386	0.237	0.094	0.018	0.311	0.244	0.098	0.018	0.354	0.226
27	521	0.135	0.020	0.387	0.297	0.110	0.018	0.310	0.298	0.110	0.019	0.355	0.257
28	499	0.045	0.018	0.467	0.059	0.037	0.017	0.343	0.059	0.035	0.018	0.370	0.046
29	354	0.019	0.018	0.510	0.002	0.007	0.014	0.371	-0.018	0.005	0.017	0.379	-0.030
30	271	0.026	0.008	0.263	0.069	0.021	0.009	0.208	0.059	0.019	0.011	0.226	0.038

These results suggest that several timers have significant ability. A majority of timers deliver average realized returns that exceed the corresponding average matched returns. While the October 1987 stock market crash generated observations that could affect inferences significantly, we have checked that the results are not sensitive to outliers associated with this event.

#### *4.2. Cumby-modest regression tests*

Cumby and Modest (1987) develop a natural extension of the nonparametric Henriksson-Merton (1981) contingency table test that is valid under more general distributional assumptions. One assumption underlying the Henriksson-Merton test is that the probability of a correct forecast is independent of the magnitude of the subsequent asset returns. Cumby and Modest argue that this assumption is unreasonable if the investment adviser is a technical analyst, which motivates them to develop an appropriate generalized test. As an alternative they propose using the regression of the equity risk premium  $R_S - R_B$  on the portfolio weight  $W$ , in which  $R_S - R_B$  and  $W$  are defined as in Section 2. If the market timer has significant ability, then the slope coefficient of the regression should be positive.

The regression test proposed by Cumby and Modest is a test for unconditional timing ability. Indeed, all tests that we have considered so far investigate unconditional, as opposed to conditional, ability. The traditional notion of timing ability is unconditional. It focuses solely on forecast accuracy, ignoring the information set used to obtain forecasts. On the other hand, conditional timing ability is forecasting ability that exists after controlling for public information available when forecasts are made. Both unconditional and conditional timing ability are of interest, and we examine both in this section.

To obtain a conditional version of the Cumby-Modest regression, we follow Ferson and Schadt (1996), Ferson and Warther (1996), and Graham and Harvey (1996). We run the original Cumby-Modest regression except that we include observable economic variables as additional regressors. Specifically, we add five lagged instruments: the yield on a one-month Treasury bill, a Treasury yield spread (ten-year minus three-month), a corporate bond yield spread (Aaa minus Baa), and the lagged return and dividend yield for the NYSE/Amex/Nasdaq portfolio. The one-month yield is based on data from the CRSP Government Bond File. Both yield spreads are based on data from the Federal Reserve Board of Governors. Both stock market variables are based on data from the CRSP Indices/Excess Returns File. The dividend yield is constructed daily, dividing total dividends for the previous 252 business days by the lagged portfolio value.

Table 3 contains unconditional and conditional regression results, obtained using autocorrelation- and heteroskedasticity-consistent standard errors as

developed by Newey and West (1987).<sup>10</sup> Consider the unconditional regressions. As in earlier tests, ability generally appears strongest relative to Nasdaq and weakest relative to the S&P 500. Using a one-sided test with a significance level of 5%, there are 17, 16, or 13 timers with significant ability based on the Nasdaq, NYSE/Amex/Nasdaq, or S&P 500 portfolios, respectively. Few estimated slope coefficients are negative, and none has a  $t$ -statistic less than  $-1$ . Timer 14 does the best, but timers 12, 23, 26, and 27 also do extremely well. Results for the conditional regressions are similar in certain respects. At least 11 timers show significant ability regardless of the portfolio used. But the conditional results differ in two major respects from results reported in our other four unconditional tests. With the conditional tests, ability is less dependent on the benchmark portfolio, and timers exhibit negative ability more frequently.<sup>11</sup>

These results demonstrate that at least some timers possess significant ability. Using either Cumby-Modest test and any portfolio, at least 11 timers always show significant ability at the 5% level. To judge whether this is noteworthy, consider the simplest scenario—timer performances are independent. Model the number of timers who show significant ability at the 5% level using a binomial  $Bin(n, p)$  distribution with  $n = 30$  trials and probability  $p = 5\%$ . Then the probability that 11 or more timers exhibit significant ability is zero, and the expected number of timers with significant ability is 1.5. Clearly, timer performances could be dependent, and hence finding 11 out of 30 timers with significant ability at the 5% level might not be as striking as this example suggests. Nonetheless, finding such ability in more than a third of our sample is impressive when compared with results previously reported.<sup>12</sup>

<sup>10</sup>In Table 3 we use all available observations. We do not delete dates such as October 19, 1987, that correspond to potentially influential observations. Given that we are testing market timing ability, deleting dates when the market moves drastically seems inappropriate. Nonetheless, we have verified that similar results hold if such dates are deleted. Details are available on request.

<sup>11</sup>These differences seem reasonable given that the conditional tests, unlike their unconditional counterparts, control for predictability based on lagged economic variables. Daily returns for the Nasdaq portfolio have higher autocorrelation than those for the other portfolios. Because Nasdaq returns are more predictable from the standpoint of autocorrelation, ability measured relative to Nasdaq could decrease more than ability measured relative to the other portfolios when switching from unconditional to conditional tests. Thus, the relatively better performance associated with Nasdaq using unconditional tests could disappear using conditional tests. In addition, controlling for predictability based on lagged economic variables could lower observed ability across all portfolios, which could lead to more negative ability being observed using conditional tests.

<sup>12</sup>For instance, Graham and Harvey (1996) investigate timing ability using an alternative regression specification. Using their Eq. (1), the percentage of timers that show significant ability at the 5% level is 8.3%. Using their Eq. (2), the percentage is 7.6%. These results are for conditional regressions; similar results hold for unconditional regressions. By way of contrast, at least 36.7% of our timers show significant ability at the 5% level regardless of the type of regression or portfolio used in Table 3.

Table 3

Cumby-Modest unconditional and conditional regression tests of market timing ability relative to the Nasdaq, NYSE/Amex/Nasdaq, and Standard & Poor's (S&P) 500 stock portfolios. All tests use the realized daily return on stocks less the realized daily return on cash as the dependent variable. These tests use the timer recommendation expressed as the portfolio weight  $W$ , i.e., the fraction of capital invested in stocks, as the independent variable of major interest. The unconditional tests include no other independent variables; the conditional tests include five predetermined economic variables used as instruments. These instruments are the yield on a one-month Treasury bill, a Treasury yield spread (ten-year minus three-month), a corporate bond yield spread (Aaa minus Baa), and the lagged return and dividend yield for the NYSE/Amex/Nasdaq portfolio. The proxy for "stocks" is either the Nasdaq, NYSE/Amex/Nasdaq, or S&P 500 stock portfolio. The proxy for "cash" is the return on a one-month Treasury bill. The regressions estimate autocorrelation- and heteroskedasticity-consistent standard errors as developed by Newey and West (1987). The estimated slope coefficients for the timer recommendation variable in the unconditional and conditional regressions are denoted by  $\beta_u$  and  $\beta_c$ , respectively. The  $t$ -statistics that correspond to testing the null hypothesis that  $\beta_u = 0$  (respectively,  $\beta_c = 0$ ) versus the alternative hypothesis that  $\beta_u > 0$  (respectively,  $\beta_c > 0$ ) are denoted by  $t(\beta_u)$  and  $t(\beta_c)$ , respectively. The number of observations used to calculate the relevant test statistics is represented by  $N$ . The sample period for each timer begins on the day of the first signal for the timer and ends on December 30, 1994.

Timer	$N$	Nasdaq				NYSE/Amex/Nasdaq				S&P 500			
		$\beta_u$	$t(\beta_u)$	$\beta_c$	$t(\beta_c)$	$\beta_u$	$t(\beta_u)$	$\beta_c$	$t(\beta_c)$	$\beta_u$	$t(\beta_u)$	$\beta_c$	$t(\beta_c)$
1	2249	0.132	2.47	0.060	1.49	0.075	1.65	-0.009	-0.24	0.059	1.26	-0.031	-0.70
2	2249	0.106	2.11	0.062	1.66	0.024	0.54	-0.006	-0.15	0.001	0.03	-0.023	-0.52
3	2249	0.147	3.32	0.060	1.38	0.098	2.65	0.027	0.68	0.086	2.21	0.019	0.43
4	2249	0.008	0.11	-0.097	-1.10	0.021	0.32	-0.060	-0.76	0.028	0.41	-0.051	-0.59
5	2209	0.131	2.83	0.032	0.74	0.083	2.11	-0.006	-0.16	0.070	1.68	-0.018	-0.40
6	2209	0.282	2.48	0.128	1.70	0.174	1.76	0.029	0.41	0.145	1.44	-0.001	-0.01
7	2167	0.136	2.82	0.056	1.32	0.085	2.09	-0.004	-0.08	0.071	1.68	-0.021	-0.38
8	2167	0.046	0.66	-0.031	-0.66	-0.008	-0.14	-0.097	-2.02	-0.027	-0.42	-0.123	-2.21
9	2167	0.080	1.42	0.032	0.74	-0.003	-0.06	-0.045	-1.01	-0.033	-0.64	-0.074	-1.47
10	1979	-0.039	-0.75	-0.042	-0.93	-0.036	-0.89	-0.040	-0.99	-0.039	-0.94	-0.045	-1.01
11	1938	0.156	3.16	0.092	2.34	0.106	2.62	0.074	1.99	0.092	2.17	0.072	1.71
12	1687	0.304	6.41	0.226	4.97	0.235	6.10	0.210	5.24	0.231	5.69	0.230	5.17
13	1687	0.062	1.31	-0.025	-0.45	0.030	0.84	-0.043	-0.94	0.018	0.49	-0.056	-1.13
14	1374	0.782	18.33	0.780	16.50	0.667	21.71	0.746	19.08	0.677	20.36	0.809	18.09
15	1290	0.141	2.52	0.062	1.25	0.103	2.52	0.063	1.68	0.093	2.21	0.068	1.67
16	1187	0.115	1.58	0.186	2.11	0.074	1.43	0.140	2.08	0.070	1.34	0.146	2.03

17	1019	0.115	1.39	0.111	1.47	0.160	2.76	0.165	3.03	0.177	3.07	0.194	3.34
18	999	0.049	0.62	0.082	1.20	-0.012	-0.22	0.006	0.11	-0.036	-0.65	-0.023	-0.42
19	978	0.033	0.55	0.150	2.02	0.011	0.29	0.083	1.66	0.006	0.15	0.069	1.29
20	959	0.206	3.90	0.140	2.92	0.122	3.32	0.091	2.53	0.100	2.62	0.085	2.13
21	853	0.102	1.81	0.068	1.25	0.021	0.55	-0.003	-0.07	0.002	0.05	-0.016	-0.40
22	853	0.113	1.20	0.054	0.90	0.013	0.31	-0.037	-0.87	-0.018	-0.43	-0.066	-1.43
23	853	0.456	6.50	0.406	5.77	0.309	6.23	0.300	6.02	0.271	5.53	0.296	5.77
24	687	-0.023	-0.27	-0.047	-0.57	-0.036	-0.66	-0.060	-1.14	-0.043	-0.82	-0.067	-1.23
25	603	0.210	2.81	0.185	2.72	0.132	2.56	0.114	2.29	0.108	2.05	0.095	1.76
26	603	0.393	7.30	0.367	7.22	0.323	8.27	0.312	8.22	0.341	8.29	0.344	8.23
27	521	0.476	7.29	0.442	6.63	0.384	8.66	0.377	8.06	0.382	8.49	0.401	7.94
28	499	0.129	1.65	0.113	1.50	0.096	1.72	0.090	1.63	0.085	1.55	0.089	1.56
29	354	0.019	0.21	0.018	0.21	-0.021	-0.33	-0.030	-0.48	-0.036	-0.56	-0.052	-0.81
30	271	0.104	1.09	-0.213	-1.06	0.072	1.06	-0.153	-0.93	0.053	0.77	-0.127	-0.75

Even if performance is dependent across timers, a simple way exists to demonstrate that our results are significant. This procedure, the Bonferroni method of multiple comparisons, allows us to test whether joint performance is significant by using information regarding individual performance reported in Table 3. (See Morrison, 1976, pp. 114–116.) Specifically, assume that the null hypothesis is that no timer has significant ability. The alternative hypothesis is that at least one timer has significant ability. To test the 30 timers simultaneously at the 0.05 level, we use the individual regression statistics reported in Table 3, except that the critical value of the test statistic now corresponds to a test of size  $0.05/30$ , not 0.05. The critical value of the  $t$ -statistics for both unconditional and conditional betas now equals 2.93, not 1.65. Regardless of the underlying portfolio, five or more timers always have  $t$ -statistics that exceed 2.93. Hence, the evidence strongly suggests that at least one timer has significant unconditional ability.<sup>13</sup>

#### 4.3. *Graham-Harvey weight change tests*

In this section we examine whether changes in market timer recommendations correctly anticipate changes in the market risk premium. The intuition is straightforward: If a timer has significant ability, then changes in the equity weight of the recommended portfolio should vary positively with changes in market returns. Our analysis parallels that of Graham and Harvey (1996), who use such an approach to study investment newsletter recommendations. (See, in particular, Table 1 of their paper.)

We construct our tests as follows. For each timer, we consider each instance in which the timer changes the recommended equity weight. Whenever this occurs, we compute average daily equity risk premiums over holding periods of 1, 5, 25, and 100 days. The holding periods begin immediately following the day that the recommendation changes.<sup>14</sup> For each holding period, we pool the cases across all timers and divide the sample into two parts corresponding to weight increases and decreases. For each of these two subsamples, we compute

<sup>13</sup>We thank Wayne Ferson for his suggestion to use the Bonferroni method. Using multivariate regression to test joint performance is an obvious alternative. A major obstacle in using regression analysis, however, is that these timers do not operate over the same time period. Therefore, one must either run the regression over the common time period, which is barely one year, or adjust for missing observations, which is nontrivial. As another alternative to testing joint performance, we could combine the forecasts of the timers into that of a consensus timer. We do this in Section 5.1.

<sup>14</sup>These timers often change recommendations at intervals shorter than 5, 25, or 100 days. As in Graham and Harvey (1996), we use shorter holding periods when appropriate. For instance, suppose we are considering a timer who is currently invested entirely in stocks. The timer switches to cash for ten days and then switches back to stocks. Then the 25-, 50-, and 100-day holding periods contain only ten days; i.e., only those days for which the new recommendation holds. Hence, the stated length for a given holding period can, and often does, exceed the actual length.





*Given last three recommendations were incorrect, equity weights*

Increased	66.7	0.24	71.4	0.18	60.0	0.02	69.2	0.06
Decreased	36.4	-0.24	50.0	-0.01	50.0	-0.17	66.7	-0.04
(p-value)	(0.089)	(0.063)	(0.131)	(0.057)	(0.274)	(0.106)	(0.433)	(0.209)

Panel C. The results below use the 15 worst timers, selected on the basis of the t-statistics of the Cumby-Modest unconditional regressions using the Nasdaq returns

*All observations in which recommended equity weights*

Increased	68.2	0.41	65.2	0.11	66.2	0.05	66.2	0.04
Decreased	43.4	-0.22	43.4	-0.07	57.1	-0.02	60.1	-0.02
(p-value)	(0.000)	(0.000)	(0.000)	(0.000)	(0.031)	(0.010)	(0.106)	(0.017)

*Given last recommendation was correct, equity weights*

Increased	66.1	0.44	65.5	0.09	68.6	0.02	70.0	0.02
Decreased	40.5	-0.25	47.6	-0.02	52.5	-0.02	58.9	-0.02
(p-value)	(0.000)	(0.000)	(0.003)	(0.013)	(0.010)	(0.195)	(0.054)	(0.198)

*Given last recommendation was incorrect, equity weights*

Increased	68.4	0.35	64.3	0.10	63.0	0.07	62.3	0.06
Decreased	50.0	-0.06	36.4	-0.12	63.2	-0.03	60.6	-0.03
(p-value)	(0.013)	(0.002)	(0.000)	(0.000)	(0.514)	(0.016)	(0.413)	(0.035)

*Given last three recommendations were correct, equity weights*

Increased	66.7	0.53	76.3	0.13	80.8	0.10	85.7	0.11
Decreased	29.5	-0.44	48.7	0.03	61.8	0.02	61.3	0.04
(p-value)	(0.000)	(0.000)	(0.006)	(0.145)	(0.056)	(0.181)	(0.029)	(0.233)

*Given last three recommendations were incorrect, equity weights*

Increased	47.1	0.09	75.0	0.02	62.5	0.00	65.4	0.06
Decreased	61.5	0.27	14.3	-0.23	53.8	-0.07	57.1	-0.12
(p-value)	(0.785)	(0.761)	(0.005)	(0.078)	(0.305)	(0.305)	(0.305)	(0.078)

two statistics: (1) the percentage of times that the average risk premium is positive, and (2) the mean of the average risk premium. If the timer has significant ability, then we expect both statistics to be greater when the timer recommends increasing, instead of decreasing, the equity weight of the recommended portfolio. Therefore, we test whether each statistic is significantly greater when weights increase than when weights decrease.

Table 4 presents these results. We conduct tests using all three indices, but to conserve space we report only the S&P 500 results. Panel A contains results using all timers. Using a 1-day horizon, the average risk premium is positive 81.9% of the time following a weight increase and 25.3% of the time following a weight decrease, a difference that is highly significant. By way of comparison, the 1-day risk premium is positive 54.0% of the time regardless of whether the timer changed weights. Using a 5-, 25-, or 100-day horizon yields similar results. The 5-, 25-, and 100-day premiums are positive 58.1%, 59.6%, and 72.7% of the time, respectively, regardless of whether the timer changed weights. Following a weight increase, the average risk premium is always positive at least 77.9% of the time. Following a weight decrease, the average risk premium is never positive more than 37.0% of the time. The differences between the corresponding percentages when weights increase instead of decrease are highly significant for all time horizons. Thus, these timers appear to be correctly anticipating changes in the market.

We also examine whether these timers exhibit “hot hands” or “cold hands.” The term “hot hands” (“cold hands”) refers to the tendency for a timer to correctly (incorrectly) anticipate the direction of the market given that the last recommendation(s) correctly (incorrectly) anticipates the direction of the market. Panel A gives results using all timers conditional on the last one or three recommendations being either correct or incorrect. Strong evidence exists of the hot hands phenomenon. Assume that the last recommendation is correct and consider all time horizons. The timers are even more accurate than they were before. Following a weight increase, the average risk premium is always positive at least 81.2% of the time. Following a weight decrease, the average risk premium is never positive more than 31.1% of the time. The differences between the corresponding percentages when weights increase instead of decrease are highly significant for all time horizons. Moreover, if the last three recommendations are correct, the timers are even more accurate. The average risk premium is always positive at least 91.1% of the time following a weight increase and is never positive more than 17.5% of the time following a weight decrease, yielding highly significant differences between the corresponding percentages once again. While there is evidence of the cold hands phenomenon, the evidence is not as strong. If the last recommendation is incorrect, the timers are less accurate than they are in the unconditional case. Nonetheless, the differences between the corresponding percentages are still highly significant. If the last three recommendations are incorrect, however,

then the differences are no longer statistically significant except for the 5-day horizon.

A natural issue is whether the results of Panel A are driven by a few good timers. The best timers tend to change their recommendations more frequently than the worst timers. For example, approximately 19% of the weight changes examined in Panel A correspond to timer 14. To explore this issue, we divided the timers into two groups, the 15 “best” and the 15 “worst,” ranking the timers by the magnitude of the  $t$ -statistic of the betas obtained in the unconditional Cumby-Modest regressions using Nasdaq returns. Panels B and C contain results for the best and worst timers, respectively. The results for the best timers are strong as would be expected given Panel A. The results for the worst timers are more surprising. They also perform relatively well in this test. Following a weight increase, the average risk premium is always positive at least 65.2% of the time. Following a weight decrease, the average risk premium is never positive more than 60.1% of the time. The differences between the corresponding percentages when weights increase instead of decrease are highly significant for all time horizons shorter than 100 days. There is also some evidence that the hot hands and cold hands phenomena hold for this subsample when the time horizon is only 1 or 5 days.

As noted earlier in footnote 14, we use a holding period shorter than 5, 25, or 100 days whenever a timer changes recommendations prior to 5, 25, or 100 days. This procedure yields returns that do not have holding periods of exactly the same length. Hence, the assumption that these returns have identical distributions is questionable. To address this issue, we repeat these tests by restricting attention to observations with holding periods that were exactly 5, 25, or 100 days. That is, if a timer changes recommendations prior to 5, 25 or 100 days, we discard the observation. This results in substantially fewer observations but solves the problem of having returns of different holding period lengths. The results, which are not reported in Table 4, are similar to those in Table 4, but somewhat weaker. The results using a 5-day horizon are still highly significant, but those using a 25- or 100-day horizon are generally insignificant. Of course, the results using a 1-day holding period are unchanged.

#### 4.4. *Nonparametric tests*

We also conduct two types of nonparametric tests to check whether the results from the previous tests are robust with respect to the methodologies used. These two types of tests are (1) Henriksson-Merton (1981) contingency table tests and (2) nonparametric measure of association tests. Based on the theoretical framework developed in Merton (1981), the contingency table tests have been used by Henriksson and Lessard (1982), Cumby and Modest (1987), and others. Our nonparametric measure of association tests, which use the

Spearman rank correlation coefficient, appear new to the market timing literature.

For the contingency table tests, consider a timer who forecasts the relative performance of stocks and cash at the start of each period. Let  $X$  equal one if the timer forecasts that stocks will beat cash and zero otherwise. Let  $Y$  equal one if stocks beat cash and zero otherwise. Given a sample of bivariate observations  $(X, Y)$ , one can classify the observations using a  $2 \times 2$  contingency table. The null hypothesis of no timing ability corresponds to independence of  $X$  and  $Y$ . Hence, to test for ability, one can test for independence in  $2 \times 2$  contingency tables. For further details regarding this test, underlying assumptions, and generalizations, see Henriksson and Merton (1981), Cumby and Modest (1987), and Pesaran and Timmermann (1994). Our situation differs slightly from those examined in Henriksson and Merton (1981) and Cumby and Modest (1987). Our timers need not recommend that a client invest entirely in one asset. Half of our timers occasionally recommend investing in both stocks and cash simultaneously. Hence, instead of using  $2 \times 2$  contingency tables exclusively, we use  $2 \times n$  contingency tables in which  $n$  equals the number of distinct portfolio weights furnished by a given timer. The test statistic is a Pearson chi-square.<sup>15</sup>

Briefly, we find that several timers possess significant ability, although the evidence is not as strong as in the previous tests. The strongest evidence of timing ability occurs when Nasdaq proxies for the market. There are eleven, seven, and six timers who exhibit significant ability relative to Nasdaq, NYSE/Amex/Nasdaq, and the S&P 500, respectively. In particular, timers 14, 23, 26, and 27 once again show significant ability across all portfolios. Detailed results are available on request.

For the nonparametric measure of association tests, we calculate the Spearman rank correlation between two variables. The first variable is the timer recommendation expressed as the portfolio weight  $W$ , in which  $W$  represents the fraction of capital invested in stocks. The second variable is an indicator variable that equals either one or zero, respectively, depending on whether the realized daily return on stocks exceeds the daily return on cash or vice versa.<sup>16</sup> If the timer has significant ability, then this correlation should be significantly positive.

To summarize, we find that over two-thirds of the estimated correlations are positive. Consistent with our previous results, the strongest evidence of timing

<sup>15</sup>These contingency table tests require an alternative hypothesis of (two-sided) association instead of positive association. We, however, are more interested in positive association than negative association. With the tests in previous sections and the association tests described later in this section, we can use one-sided alternative hypotheses, and hence we do so.

<sup>16</sup>We also run an alternative test using the equity risk premium instead of the corresponding indicator variable. The results are qualitatively similar.

ability occurs using Nasdaq. In that case, 17 timers show significant ability based on one-sided statistical tests at the 5% significance level. The corresponding numbers are ten and seven for the NYSE/Amex/Nasdaq and S&P 500 portfolios, respectively. Timer 14 turns in the most impressive performance with estimated correlations of 0.428, 0.416, and 0.350 for the three indices. Timers 12, 23, 25, 26, and 27 also exhibit significant ability regardless of which stock portfolio is used. Detailed results are available on request.

## 5. Related issues in testing performance

The tests in Section 4 demonstrate that the timers appear to have significant ability. Several secondary issues, however, still remain, such as consensus and statistical forecasting, nonsynchronous trading, survivorship bias, persistence, transaction costs and management fees, the frequency with which a timer changes recommendations, and the frequency with which a researcher observes recommendations.

### 5.1. *Consensus and statistical timers*

Following Graham and Harvey (1994), we examine two hypothetical timers. We first construct the forecasts of an equally weighted timer, which we shall refer to as a consensus timer. Starting on the first day of the data set, we allocate money equally among all timers who had data available for the next day. We then compute the equally weighted average return of the timers over the next day. This continues on a daily basis, with an equal allocation, until a new timer enters or an old one leaves. When that happens, we simply add or remove a timer, continuing to allocate money equally across all timers with data available. We then evaluate the performance of this consensus timer using several of the tests discussed in Section 4. The results are presented in Panel A of Table 5.

The consensus timer performs exceptionally well, particularly with respect to Nasdaq. If the consensus timer can be used as an indicator of the overall timing ability of this set of timers, then we would conclude that these timers have statistically significant ability.

In Section 4.2, we discussed conditional tests in which macroeconomic and financial variables were added to our Cumby-Modest regressions to help identify whether these timers were demonstrating ability beyond what could be obtained by simply using contemporaneous information. Following Graham and Harvey (1994), we use this format to construct another hypothetical timer, referred to as a statistical timer, who forecasts using a naive model. Specifically, we estimate a regression of the market risk premium on the conditioning

variables of Section 4.2. For a given month, if the regression forecasts a positive risk premium, we assume the timer invests entirely in stocks. If the regression forecasts a negative risk premium, we assume that the timer invests entirely in cash. We then evaluate this statistical timer using the same tests as in Section 4. The results are shown in Panel B of Table 5.

The statistical timer performs well in the ratio and unconditional tests, but not in the conditional tests. Those results are not significant, but we would not expect them to be, inasmuch as that test simply shows whether there is information in the forecasts beyond that provided by the conditioning economic and financial variables. Because those variables are used to construct the forecast, they can provide no additional information of any significance.

Because the statistical timer demonstrates significant ability, we might conclude that successful market timing can be accomplished using simple information freely available. But we have also found that the timers do provide information beyond that provided by the statistical timer. So in other words, it

Table 5  
 Tests based on two hypothetical timers, one that uses consensus forecasts of all timers and one that uses forecasts constructed using a simple statistical model. The proxy for “stocks” is either the Nasdaq, NYSE/Amex/Nasdaq, or Standard & Poor’s (S&P) 500 stock portfolio. The proxy for “cash” is the return on a one-month Treasury bill. As in Table 2, the performance measure Ratio is the difference between Sharpe ratios for the timer’s portfolio and the corresponding volatility-matched portfolio. As in Table 3, the estimated slope coefficients for the timer recommendation variable in the unconditional and conditional regressions are denoted by  $\beta_u$  and  $\beta_c$ , respectively. The  $t$ -statistics that correspond to testing the null hypothesis that  $\beta_u = 0$  (respectively,  $\beta_c = 0$ ) versus the alternative hypothesis that  $\beta_u > 0$  (respectively,  $\beta_c > 0$ ) are denoted by  $t(\beta_u)$  and  $t(\beta_c)$ , respectively. The sample period for each timer begins on the day of the first signal for the timer and ends on December 30, 1994.

Performance measure	Nasdaq	NYSE/Amex/Nasdaq	S&P 500
<i>Panel A. Consensus timer constructed as an equally weighted average return over all timers for which data are available from the period 1/2/86 through 12/31/94</i>			
Ratio	0.08	0.05	0.04
$t(\beta_u)$	5.26	4.04	3.47
$t(\beta_c)$	5.11	2.86	2.17
<i>Panel B. Statistical timer constructed by estimating a regression of the market risk premium on the following conditional variables: the yield on a one-month Treasury bill, a Treasury yield spread (ten-year minus three-month), a corporate bond yield spread (Aaa minus Baa), and the lagged return and lagged dividend yield for the NYSE/Amex/Nasdaq portfolio. If the risk premium is estimated to be positive, a position of 100% in stock is taken; if the risk premium is estimated to be negative, a position of 100% in cash is taken</i>			
Ratio	0.12	0.08	0.05
$t(\beta_u)$	6.71	4.85	3.40
$t(\beta_c)$	-0.86	-0.25	0.05

may be possible to time the market using statistical rules, but based on our results so far, the skills of these timers can also add value.

### 5.2. *Nonsynchronous trading and its potential impact*

Index returns can display significant autocorrelation as a result of nonsynchronous trading of the component stocks used in the index. An index quote is a composite of prices, some of which can be current and some of which can be stale. Boudoukh et al. (1994) find that although index returns on small capitalization portfolios display significant autocorrelation, returns on the corresponding futures contract display almost none. With large capitalization portfolios, they find little autocorrelation for either index or index futures returns.

We consider two questions regarding the potential impact of nonsynchronous trading on our analysis. First, is the evidence of significant ability presented in Section 4 spurious, reflecting nonsynchronous trading and nonexploitable return autocorrelation? Second, can nonsynchronous trading explain our finding that ability appears strongest measured relative to Nasdaq and weakest measured relative to the S&P 500?

Ideally, we would repeat the tests of Section 4 using Nasdaq futures returns, not Nasdaq returns themselves. During the 1986–1994 period, however, Nasdaq futures were not actively traded. We, therefore, concentrate on the S&P 500 futures contract, which was actively traded during this period, and repeat our tests using S&P 500 futures returns instead of S&P 500 index returns. If we find a significant drop in ability using futures returns instead of index returns, then nonsynchronous trading could be a problem.

Table 6 presents results for the tests of Sections 4.1 and 4.2 using both S&P 500 futures and index returns. For futures returns we use the nearby futures contract, switching to the next contract on the first business day of the expiration month. There is little, if any, evidence that ability decreases using futures returns instead of index returns. Consider the regression results, for instance. For the unconditional regressions, either 12 or 11 timers show significant ability depending on whether we use index or futures returns. For the conditional regressions, 11 timers show significant ability regardless of whether we use index or futures returns. Thus, for the S&P 500, we find significant ability using either index or futures returns. This result, coupled with the fact that timing ability appears weakest for the S&P 500, implies that nonsynchronous trading cannot explain our finding that significant ability exists.

Can nonsynchronous trading explain our finding that ability appears strongest measured relative to Nasdaq and weakest measured relative to the S&P 500? That is certainly a possibility. Nonsynchronous trading could lead to autocorrelation that is significantly greater in Nasdaq returns than S&P 500

Table 6

Market timing performance using the Standard & Poor's (S&P) 500 nearby futures returns as the benchmark portfolio returns. Ratio is the excess return on the timer portfolio minus the excess return on the portfolio consisting of the market index and cash with the same standard deviation as the timer portfolio.  $t(\beta_u)$  is the  $t$ -statistic from a regression of the market risk premium on the timer weight.  $t(\beta_c)$  is the  $t$ -statistic from the conditional regression of the market risk premium on the timer weight and four conditioning variables: the one-month Treasury bill rate, a Treasury yield spread, a corporate bond yield spread, and the lagged return and dividend yield on the NYSE/Amex/Nasdaq portfolio. Although reported earlier in Tables 2 and 3, the corresponding statistics using the S&P index returns appear in parentheses for comparison purposes.

	Ratio		$t(\beta_u)$		$t(\beta_c)$	
Timer	Futures return	(Index return)	Futures return	(Index return)	Futures return	(Index return)
1	0.016	(0.017)	1.00	(1.26)	-0.87	(-0.70)
2	-0.003	(-0.002)	-0.22	(0.03)	-0.63	(-0.52)
3	0.021	(0.022)	2.11	(2.21)	0.51	(0.43)
4	-0.001	(-0.016)	0.45	(0.41)	-0.36	(-0.59)
5	0.015	(0.015)	1.19	(1.68)	-0.51	(-0.40)
6	0.018	(0.021)	1.20	(1.44)	-0.21	(-0.01)
7	0.023	(0.019)	1.51	(1.68)	-0.50	(-0.38)
8	-0.012	(-0.011)	-0.69	(-0.42)	-2.37	(-2.21)
9	-0.013	(-0.012)	-0.75	(-0.64)	-1.57	(-1.47)
10	-0.016	(-0.018)	-1.10	(-0.94)	-1.14	(-1.01)
11	0.030	(0.031)	1.73	(2.17)	1.39	(1.71)
12	0.083	(0.087)	5.30	(5.69)	4.97	(5.17)
13	0.004	(-0.002)	0.44	(0.49)	-0.97	(-1.13)
14	0.319	(0.346)	18.85	(20.36)	16.06	(18.09)
15	0.042	(0.039)	2.20	(2.21)	1.71	(1.67)
16	0.022	(0.018)	1.29	(1.34)	2.01	(2.03)
17	0.057	(0.054)	3.06	(3.07)	3.35	(3.34)
18	-0.032	(-0.048)	-0.72	(-0.65)	-0.46	(-0.42)
19	-0.002	(-0.006)	0.06	(0.15)	1.21	(1.29)
20	0.053	(0.046)	2.42	(2.62)	2.08	(2.13)
21	-0.006	(-0.006)	-0.24	(0.05)	-0.60	(-0.40)
22	-0.015	(-0.015)	-0.65	(-0.43)	-1.62	(-1.43)
23	0.077	(0.082)	4.98	(5.53)	5.35	(5.77)
24	-0.018	(-0.019)	-0.90	(-0.82)	-1.19	(-1.23)
25	0.065	(0.058)	1.95	(2.05)	1.71	(1.76)
26	0.231	(0.226)	8.56	(8.29)	8.52	(8.23)
27	0.242	(0.257)	7.76	(8.49)	7.46	(7.94)
28	0.058	(0.046)	1.62	(1.55)	1.69	(1.56)
29	-0.016	(-0.030)	-0.55	(-0.56)	-0.84	(-0.81)
30	0.051	(0.038)	0.80	(0.77)	-0.65	(-0.75)

returns. The fact that Nasdaq returns exhibit higher autocorrelation, and hence greater apparent predictability, is definitely consistent with our empirical results. During 1986–1994 the first-order autocorrelation of daily returns for Nasdaq, NYSE/Amex/Nasdaq, and the S&P 500 is 0.256, 0.117, and 0.046,

respectively. Unlike our unconditional tests, the conditional tests control for predictability based on lagged economic variables. This result suggests that ability could appear stronger measured relative to Nasdaq than the S&P 500 in the unconditional tests, but not the conditional tests. That is exactly what we find in Sections 4.1 and 4.2.

### 5.3. *Survivorship bias*

We do not have timing recommendations for managers who dropped out of the MoniResearch database. We do know, however, that ten managers dropped out of the database, all during the 1991–1994 period. Three timers dropped out to become asset allocators; i.e., they began recommending more than two asset classes. Three timers sold their businesses. One timer moved all clients into a money market fund and closed his business. One timer gradually moved from the equity market to the debt market, and MoniResearch decided that he did not belong in this group of equity timers. One timer moved from market timing into more traditional money management. Finally, one timer dropped out of the database with no reason given.

Given the preceding explanations, it is unlikely that all ten timers dropped out because of poor performance. Suppose, however, that they did. Assuming an original sample of 40 timers and using Cumby-Modest regression tests, at least 11 out of 40 timers have significant performance at the 5% level regardless of the test and portfolio used. Assuming performance is independent across timers and using a binomial model as in Section 4.2, then the probability that at least 11 timers demonstrate such ability is zero and the expected number of timers with such ability is two. Performance certainly could be dependent across timers. Timers using similar styles and data should produce similar performance. Without specific information regarding timer dependence, however, a binomial model seems the most reasonable (indeed, the only) model available. On the other hand, suppose timer performance is dependent across timers. One can still argue, as done in Section 4.2 using the Bonferroni method, that survivorship bias does not drive our results.

### 5.4. *Signal frequency*

Our results indicate that timers 14, 23, 26, and 27 are among the very best performers. They exhibit timing ability across all four tests and three benchmark portfolios. They also change recommendations more frequently than most timers in our sample. To investigate whether signal frequency (the frequency with which a timer changes recommendations) is associated with relative performance, we rank all timers in terms of these two variables and then calculate the correlation between rankings. Using any portfolio and either regression test, for example, the correlation ranges from 0.49 to 0.52. This

result is not surprising. Technical trading rules often do well, ignoring transaction costs, if trading occurs frequently. In that case, the impact of transaction costs becomes more critical. A trading strategy that looks phenomenal before incorporating transaction costs could look terrible after incorporating transaction costs.

### 5.5. Transaction costs and management fees

Given the evidence of market timing ability in our previous sections, we raise the hurdle by redefining our tests of Section 4.1 to incorporate a cost per transaction for each weight change and a management fee. Based on information we gathered from reliable sources in the industry, we test transaction fees ranging from 0.0% to 0.5% per switch and management fees of 0.0% to 2.0%.<sup>17</sup> To conserve space we do not show these results in tabular form, but they are available on request.

Consider timers 14, 23, 26 and 27, who perform so well in Tables 2 and 3. Timers 14 and 23 earn positive excess returns regardless of the transaction costs and management fee assumed. The results for timers 26 and 27, however, are more sensitive to transaction costs. If transaction costs are 0.5%, these timers earn negative excess returns. If transaction costs are 0.25%, they earn excess returns of 10% or more. The number of timers having nonnegative excess returns ranges from a high of 26 to a low of 15 depending on the transaction costs and management fee assumed. Management fees do have an impact, but because successful timers frequently switch between asset classes, management fees are secondary in importance to transaction costs.

### 5.6. Persistence

We next investigate whether market timing ability persists over time. Relevant evidence regarding market timers has been given by Graham and Harvey (1996), who conclude that poor performance is far more persistent than good performance. To gauge persistence in our timer sample, we again use nonparametric measures of association. For a given performance measure and stock portfolio, we compute Spearman correlation coefficients between rankings of market timers across two subperiods. This yields a relative, as opposed to absolute, measure of persistence. That is, we examine whether

<sup>17</sup>Transaction costs were estimated using the equation  $(252/N)[N\mu(R) - ST] - F - 252\mu(M)$ , in which  $N$  = number of daily observations for the timer,  $S$  = number of signals given by the timer,  $T$  = the transaction cost charged for changing weights,  $\mu(R)$  = mean daily return realized by the timer,  $\mu(M)$  = mean daily return for volatility-matched portfolio, and  $F$  = annual management fee. The intuition is that over the period, a timer has a nonannualized mean return of  $N\mu(R)$  before transaction costs and  $N\mu(R) - ST$  after transaction costs. To annualize, we multiply  $N\mu(R) - ST$  by  $252/N$ . We then subtract  $F$  and  $252\mu(M)$  to obtain the excess return.

timers who outperform (respectively, underperform) their peers during one subperiod perform similarly during the following subperiod. The subperiods for a given timer consist of the first and second halves of the entire sample for which the timer gives recommendations.

All 30 timers are used in calculating the correlation across subperiods for the first performance measure (i.e.,  $[\mu(R) - \mu(M)]/\sigma$ ). Timer 30, however, is not used in calculating the other three correlations. Because timer 30 makes only one recommendation in the second subperiod, one cannot calculate the last four performance measures for that subperiod. The  $t$ -statistics associated with the estimated correlations correspond to a  $t$ -distribution having either 28 or 27 degrees of freedom, depending on whether timer 30 is included or excluded, respectively.<sup>18</sup>

Table 7 presents the desired Spearman rank correlations and associated  $t$ -statistics. Evidence supporting persistence is strongest relative to the unconditional Cumby-Modest regression tests. For this test, the correlations always exceed 0.4 and always are significant at the 5% level regardless of the portfolio used. The evidence is not as strong for the ratio test of Table 2, but it is significant when the Nasdaq portfolio is used. With the conditional Cumby-Modest regressions, the situation is reversed; persistence is significant when portfolios other than Nasdaq are used.

These results suggest that relative timing ability persists over time. Admittedly, the evidence is weaker with respect to the first performance measure than the second measure. But the first performance measure is fairly arbitrary, and the significance level of the corresponding test is unknown. Therefore, the fact that evidence appears weaker for the first measure is not surprising.

To gain further insight, we also analyze persistence by following the timers on an annual basis. The methodology is as follows. First, we divide the timers into cohorts from 1986 through 1994 based on whether they were giving recommendations at the start of a given year. For example, the 1988 cohort consists of timers 1–11. Next, for each cohort we divide the timers into two groups, superior and inferior timers, based on their performance, as measured by  $t$ -statistics from the unconditional Nasdaq tests in Section 4.2, during the year that determines the cohort. With the 1988 cohort, for instance, the superior timers are the six that do the best during 1988 and the inferior timers are the five that do the worst in 1988. Finally, for each cohort we compute the difference between the average  $t$ -statistics of the two groups annually. We track these results year by year until the end of the data set. To conserve space, we do not show figures or tables but the results are available on request.

<sup>18</sup> We calculate  $t$ -statistics using the fact  $\rho\sqrt{(n-2)/(1-\rho^2)}$  is approximately distributed as a  $t$ -distribution with  $n-2$  degrees of freedom. See Kendall and Stuart (1979, p. 503) for details.

Table 7

Tests for persistence of relative performance using Spearman rank correlation coefficients between relative rankings of market timers across subperiods. The subperiods for a given timer consist of the first and second halves of the entire sample for which the timer supplies recommendations. There are three performance measures and three stock indices. The three performance measures are the Ratio used in Table 2 and the  $t$ -statistics  $t(\beta_u)$  and  $t(\beta_c)$  used in Table 3. Timers are ranked according to each measure. The three stock indices correspond to the Nasdaq, NYSE/Amex/Nasdaq, and Standard & Poor's (S&P) 500 stock portfolios, respectively. All 30 timers are used in calculating the correlation across subperiods for the first performance measure. Timer 30 is not used in calculating the other correlations. (Because timer 30 makes only one recommendation in the second subperiod, one cannot calculate the other performance measures for that subperiod.) The  $t$ -statistics associated with the estimated correlations correspond to a  $t$ -distribution having either 28 or 27 degrees of freedom, depending on whether timer 30 was included or excluded, respectively. The corresponding critical values for one-sided tests of no association versus positive association are 1.701 and 1.703, respectively, where the level of the tests is 0.05.

	Nasdaq	NYSE/Amex/Nasdaq	S&P 500
Performance measure	Correlation ( $t$ -statistic)	Correlation ( $t$ -statistic)	Correlation ( $t$ -statistic)
Ratio	0.381 (2.181)	0.272 (1.496)	0.340 (1.911)
$t(\beta_u)$	0.485 (2.932)	0.441 (2.599)	0.508 (3.120)
$t(\beta_c)$	0.174 (0.937)	0.321 (1.795)	0.348 (1.963)

Although there is a general tendency for the series to decrease over time, 75% of the observations are positive. The later cohorts tend to outperform the earlier cohorts. This could reflect the fact that they cover shorter time periods, and so the timers are not monitored as long. It could reflect time varying performance by the stock and bond markets in the first and second halves during 1986–1994. It could also reflect the fact that as time passes, the cohorts contain more timers, leading to average  $t$ -statistics based on increasingly larger samples.

The behavior of our four best timers – 14, 23, 26, and 27 – is worth examining in detail. Their performances are striking because of their strength and consistency. Each timer demonstrates significant ability every year.

5.7. *The frequency with which recommendations are observed*

Timers 14, 23, 26, and 27 show ability across all tests and portfolios. They also change their recommendations relatively often. Therefore, we consider the following question: Do these timers still perform well if their recommendations are held constant each month? That is, assume that the recommendation for the first business day of a month remains the same during the remainder of that month. Do such timers still exhibit significant ability?

In a sense, this question asks whether detection of timing ability depends on the frequency with which a timer's recommendation is observed. Section 5.4 suggests that successful timers tend to switch recommendations more often. Suppose that a timer truly has ability and frequently switches between stocks and bonds. If that timer's recommendations are observed daily, then all changes in recommendations are observed. But suppose a timer's recommendations are observed less frequently, e.g., monthly, and then assumed to be constant over that length of time. This could lead to misrepresentation of the recommendations, which in turn could lead to underestimation of the timer's true ability.<sup>19</sup>

To determine whether such underestimation can occur, we test the performance of four hypothetical timers. These hypothetical timers, which we call newtimers 14, 23, 26, and 27, correspond to timers 14, 23, 26, and 27, respectively. Each newtimer gives the same recommendation as the corresponding original timer on the first business day of each month, but then keeps that recommendation constant for the rest of the month.

Table 8 provides strong evidence that the newtimers do not inherit the significant ability displayed by their original counterparts. Using the mean-standard deviation test of Section 4.1, there is some evidence of ability. Timers 14 and 27 generate positive performance measures for all three portfolios, for instance. But we cannot say that these values are positive at a significance level of 5%. Hence, consider the other two tests, which can be run at a significance level of 5%. Because there are four timers, two tests, and three portfolios, there are 24 cases. Out of these 24 cases, only once does a newtimer exhibit significant ability at the 5% level.<sup>20</sup> That case occurs with newtimer 27 using the conditional Cumby-Modest test and the S&P 500. These results tell a completely different story from the corresponding results for the original timers. They demonstrate that the frequency with which recommendations are observed can play a crucial role in detecting ability.

## 6. Conclusions

We provide new evidence regarding the ability of professional market timers. A notable feature of the analysis is its relatively unique data; i.e., timer

<sup>19</sup>Goetzmann et al. (2000) and Bollen and Busse (2001) both conclude that the effectiveness of market timing can be understated when monthly data are used for timers who make recommendations on a daily basis. Similarly, in a paper involving mutual fund tournaments, Busse (2001) finds that inferences regarding managerial performance can hinge on whether one examines daily or monthly returns.

<sup>20</sup>If we also include the analogous results from the nonparametric tests in Section 4.4, only once in 60 cases does a newtimer display significant ability at the 5% level.

Table 8

Tests for timing ability using hypothetical timers whose timing recommendations are constant throughout each month but agree with the recommendations of the original timers at the start of each month. Define recommendations for four new timers, henceforth called newtimers 14, 23, 26, and 27, as follows: Each newtimer gives recommendations that are constant throughout a given month but agree with the recommendation of the corresponding timer on the first business day of that month. The tests use three performance measures and three stock indices. The three performance measures are the Ratio used in Table 2 and the  $t$ -statistics  $t(\beta_u)$  and  $t(\beta_c)$  used in Table 3. The three stock indices correspond to the Nasdaq, NYSE/Amex/Nasdaq, and Standard & Poor's (S&P) 500 stock portfolios, respectively.

Newtimer	Ratio	$t(\beta_u)$	$t(\beta_c)$
<i>Panel A. Nasdaq</i>			
Newtimer 14	0.016	0.76	0.07
Newtimer 23	0.006	0.72	-0.79
Newtimer 26	-0.029	-0.25	-1.47
Newtimer 27	0.016	0.61	0.61
<i>Panel B. NYSE/Amex/Nasdaq</i>			
Newtimer 14	0.007	0.51	-0.25
Newtimer 23	-0.001	0.08	-1.41
Newtimer 26	-0.041	-0.85	-2.04
Newtimer 27	0.018	0.72	1.27
<i>Panel C. S&amp;P 500</i>			
Newtimer 14	0.004	0.36	-0.32
Newtimer 23	-0.004	-0.23	-1.51
Newtimer 26	-0.043	-1.08	-2.09
Newtimer 27	0.021	0.93	1.67

recommendations that are explicitly known and executed in customer accounts. Using four types of tests and three benchmark portfolios, we examine both unconditional and conditional timing ability on a daily basis. Contrary to most prior research, we find evidence of significant ability across all tests and portfolios. Transaction costs and survivorship bias reduce, but do not eliminate, evidence of significant ability. Relative performance persists over time and varies with the frequency with which a timer changes recommendations. We construct hypothetical consensus and statistical timers and evaluate their performance. The consensus timer displays both conditional and unconditional ability. Because the statistical timer forecasts based on certain macroeconomic variables, we expect it to show only unconditional ability, and that is the case. When recommendations of successful timers are observed monthly instead of daily, significant ability generally disappears. Thus, the frequency with which recommendations are observed can seriously affect inferences regarding timing ability.

Although we find evidence of ability across all portfolios, the evidence generally appears strongest relative to Nasdaq and weakest relative to the S&P 500. We can only conjecture as to why this pattern occurs. We know that these timers usually invest in small capitalization or growth funds. If timers provide recommendations tailored for a fund that is more similar to Nasdaq than the S&P 500 in terms of its composition, then it is not surprising that the recommendations fare better relative to Nasdaq than the S&P 500. For example, certain timers might use a forecasting model that incorporates the relatively high autocorrelation exhibited by Nasdaq returns, in which case their performance should look better relative to Nasdaq than the S&P 500. However, the apparently better performance obtained using Nasdaq returns might be spurious, in the sense that it could reflect nonsynchronous trading and nonexploitable return autocorrelation.

In closing, we believe three factors potentially explain why this study, unlike most of its predecessors, finds evidence of timing ability. First, the timers studied here are professionals who execute their recommendations for clients. If any timers possess significant ability, these timers are likely candidates. Second, because the recommendations studied here are explicitly known, they are free of estimation error, unlike implicit recommendations estimated from mutual fund returns. Third, the recommendations studied here are observed daily, not monthly or quarterly. Daily observation could aid detection of significant ability, especially for timers who frequently switch between asset classes. However, nonsynchronous trading does not explain our finding of timing ability. We know that timing ability appears weakest for the S&P 500, and for the S&P 500 nonsynchronous trading does not lead to spurious evidence of ability. Consequently, nonsynchronous trading cannot explain our conclusion that significant timing ability exists.

## References

- Admati, A., Bhattacharya, S., Pfleiderer, P., Ross, S., 1986. On timing and selectivity. *The Journal of Finance* 41, 715–730.
- Alexander, G., Benson, P., Eger, C., 1982. Timing decisions and the behavior of mutual fund systematic risk. *Journal of Financial and Quantitative Analysis* 17, 579–602.
- Bollen, N., Busse, J.A., 2001. On the timing ability of mutual fund managers. *The Journal of Finance* 56, 1075–1094.
- Boudoukh, J., Richardson, M., Whitelaw, R., 1994. A tale of three schools: insights on autocorrelation of short-horizon stock returns. *The Review of Financial Studies* 7, 539–573.
- Brocato, J., Chandy, P., 1994. Does market timing really work in the real world? *The Journal of Portfolio Management* 20 (Winter), 39–44.
- Busse, J., 2001. Another look at mutual fund tournaments. *Journal of Financial and Quantitative Analysis* 36, 53–73.
- Chang, E., Lewellen, W., 1984. Market timing and mutual fund investment performance. *The Journal of Business* 57, 57–72.

- Cumby, R., Modest, D., 1987. Testing for market timing ability: a framework for forecast valuation. *Journal of Financial Economics* 19, 169–189.
- Dybvig, P., Ross, S., 1985. Differential information and performance measurement using a security market line. *The Journal of Finance* 40, 383–399.
- Elton, E., Gruber, M., 1991. Differential information and timing ability. *Journal of Banking and Finance* 15, 117–131.
- Ferson, W., Schadt, R., 1996. Measuring fund strategy and performance in changing economic conditions. *The Journal of Finance* 51, 425–462.
- Ferson, W., Warther, V., 1996. Evaluating fund performance in a dynamic market. *Financial Analysts Journal* 52 (November/December), 20–28.
- Goetzmann, W., Jorion, P., 1999. Re-emerging markets. *Journal of Financial and Quantitative Analysis* 34, 1–32.
- Goetzmann, W., Ingersoll, J., Ivkovich, Z., 2000. Monthly measurement of daily timers. *Journal of Financial and Quantitative Analysis* 3, 257–290.
- Graham, J., 1999. Herding among investment newsletters: theory and evidence. *The Journal of Finance* 54, 237–268.
- Graham, J., Harvey, C., 1994. Market timing ability and volatility implied in investment newsletters' asset allocation recommendations. Unpublished working paper. National Bureau of Economic Research, Cambridge, MA.
- Graham, J., Harvey, C., 1996. Market timing ability and volatility implied in investment newsletters' asset allocation recommendations. *Journal of Financial Economics* 42, 397–422.
- Graham, J., Harvey, C., 1997. Grading the performance of market-timing newsletters. *Financial Analysts Journal* 53 (November/December), 54–66.
- Grinblatt, M., Titman, S., 1989. Portfolio performance evaluation: old issues and new insights. *The Review of Financial Studies* 2, 393–421.
- Grinblatt, M., Titman, S., 1994. A study of monthly mutual fund returns and performance evaluation techniques. *Journal of Financial and Quantitative Analysis* 29, 419–444.
- Henriksson, R., Lessard, D., 1982. The efficiency of the forward exchange market: a conditional nonparametric test of forecasting ability. Unpublished working paper. University of California, Berkeley.
- Henriksson, R., Merton, R., 1981. On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills. *The Journal of Business* 54, 513–533.
- Jaffe, J., Mahoney, J., 1999. The performance of investment newsletters. *Journal of Financial Economics* 53, 289–307.
- Kendall, M., Stuart, A., 1979. *The Advanced Theory of Statistics*. Vol. 2: Inference and Relationship, 4th Edition, Macmillan, New York.
- Kester, G., 1990. Market timing with small versus large-firm stocks: potential gains and required predictive ability. *Financial Analysts Journal* 46 (September/October), 63–69.
- Kon, S., 1983. The market-timing performance of mutual fund managers. *The Journal of Business* 56, 323–347.
- Kon, S., Jen, F., 1978. Estimation of time-varying systematic risk and performance for mutual fund portfolios. *The Journal of Finance* 33, 457–475.
- Kon, S., Jen, F., 1979. The investment performance of mutual funds: an empirical investigation of timing, selectivity, and market efficiency. *The Journal of Business* 52, 263–290.
- Lee, C., Rahman, S., 1990. Market timing and mutual fund performance: an empirical investigation. *The Journal of Business* 63, 261–278.
- Lehmann, B., Modest, D., 1987. Mutual fund performance evaluation: a comparison of benchmarks and benchmark comparison. *The Journal of Finance* 42, 233–265.
- Merton, R., 1981. On market timing and investment performance. I. An equilibrium theory of value for market forecasts. *The Journal of Business* 54, 363–406.

- Metrick, A., 1999. Performance evaluation with transactions data: the stock selection of investment newsletters. *The Journal of Finance* 54, 1743–1775.
- Morrison, D., 1976. *Multivariate Statistical Methods*. McGraw-Hill, New York.
- Newey, W., West, K., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.
- Pesaran, M., Timmermann, A., 1994. A generalization of the non-parametric Henriksson-Merton test of market timing. *Economics Letters* 44, 1–7.
- Treynor, J., Mazuy, K., 1966. Can mutual funds outguess the market? *Harvard Business Review* 44, 131–136.
- Wagner, J., Shellans, S., Paul, R., 1992. Market timing works where it matters most...in the real world. *The Journal of Portfolio Management* 18 (Summer), 86–90.