NBER WORKING PAPER SERIES

PERFORMANCE EVALUATION WITH TRANSACTIONS DATA: THE STOCK SELECTION OF INVESTMENT NEWSLETTERS

Andrew Metrick

Working Paper 6648
http://www.nber.org/papers/w6648

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 1998

I thank John Campbell, Gary Chamberlain, Paul Gompers, John Graham, Campbell Harvey, Leslie Jeng, Susie Metrick, Jack Porter, Andrei Shleifer, Sheridan Titman, Richard Zeckhauser, seminar participants at the NBER, the 1997 NBER Summer Institute, Harvard, MIT, and Chicago for useful comments. Daniel Chen and Alex Tsai for editorial assistance, and James Choi and Tom Knox for excellent and tireless research assistance. Mark Hulbert was very generous with his time, advice, and data. This project was partially supported by the Richard Hernstein Fund at Harvard University. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Performance Evaluation with Transactions Data:
The Stock Selection of Investment Newsletters
Andrew Metrick
NBER Working Paper No. 6648
July 1998
JEL No. G12, G14, G20

ABSTRACT

This paper analyzes the equity-portfolio recommendations made by investment newsletters. The dataset spans 17 years, is free of survivor and back-fill biases, and includes the complete recommendations for 153 different newsletters. Overall, there is no significant evidence of superior stock-picking ability for this sample of newsletters. Some individual letters do have superior performance records, but this does not occur more often than would be expected by chance, and these records are never more extreme than would be expected for the sample size. In addition, a strategy of buying past winners does not earn positive abnormal returns. The comprehensive and bias-free transactions database also allows for insights into several popular models of performance evaluation. The transactions-based approach of Daniel, Grinblatt, Titman and Wermers (1997) yields a median improvement in precision of 10 percent over the 4-factor model of Carhart (1997a), with the former approach providing more precise estimates of abnormal performance for more than 80 percent of the newsletters. This compares with a median improvement of less than 1 percent for the 4-factor model over the CAPM.

Andrew Metrick
Department of Economics
Harvard Univeristy
Littauer Center, North Yard
Cambridge, MA 02138
and NBER
ametrick@harvard.edu

1. Introduction

Investment newsletters have been part of the financial landscape since at least the early part of this century, and the current industry of over 500 active letters has about 2 million subscribers.[1] The typical newsletter is produced by a small staff and provides a wide range of advice targeted at the retail investor. Successful publications can earn millions of dollars in subscription revenue and allow their editors to become frequent speakers at investment seminars.[2] This paper undertakes a detailed study of newsletters' equity recommendations. The dataset spans 17 years, is free of survivor and back-fill biases, and contains every recommended (long) transaction for 153 newsletters. These data allow for an analysis of two questions. First, do investment newsletters have stock-selection ability? Second, can transactions data be used to improve the precision of performance evaluation?

The Hulbert Financial Digest (HFD) has been tracking investment-newsletter recommendations since 1980. The only previous papers to use the HFD database are Graham and Harvey (1996 and 1997) and Graham (1998), which focus on newsletters' timing between stocks and cash. The present paper is the first attempt to study the specific equity recommendations of these financial advisors.[3] In this respect, the paper is similar to other studies of "expert" equity recommendations such as Barber and Loeffler (1993) (*The Wall Street Journal*'s Dartboard column), Desai and Jain (1996) ("Superstar" money managers in *Barron's*), and Womack (1996) (brokerage analysts). To carry out the analysis, the paper

[1] Hulbert (1996).

[2] Brimelow (1986).

[3] One newsletter contained in the HFD sample, *The Value Line Investment Survey*, has received substantial attention in the academic literature. A series of papers since the 1960s has alternatively identified and explained the "Value Line Anomaly", the apparent superior performance of Value Line's recommendations. See Shelton (1967), Black (1973), Copeland and Mayers (1982), Stickel (1985), Huberman and Kandel (1990), and Lewis, Rogalski and Seward (1997) for some key papers in this debate.

uses performance-evaluation methods developed in several recent studies of mutual funds (Malkiel (1995), Carhart (1997a), and Daniel, Grinblatt, Titman and Wermers [DGTW] (1997)). Using these methods, I find no evidence of abnormal stock-selection ability in this sample of newsletters.

In addition to determining the investment value of newsletters' stock selection advice, this paper also attempts to shed light on the power and limitations of various performance-evaluation methodologies. The transactions-level detail and bias-free construction of the HFD database allow for a comparison of results from several different methods and insight into their relative precisions. While such comparisons can also be carried out on simulated data, such tests may miss crucial elements of actual managed portfolios and can lead to biased results. The HFD database provides a rare "natural" experiment. I find that the transactions-based approach of DGTW (1997) yields a median improvement in precision of 10 percent over the 4-factor model of Carhart (1997a), with the former approach providing more precise estimates of abnormal performance for over 80 percent of the newsletters. This compares with a median improvement of less than 1 percent for the 4-factor model over the CAPM. These results have useful implications given the advent of transactions and "holdings" databases in studies of insider trading (Eckbo and Smith (1998), Jeng (1998)), individual investors (Barber and Odean (1998), Grinblatt and Keloharju (1998)), and mutual funds (DGTW (1997)).

Section 2 of the paper describes the HFD database in detail and provides summary statistics on newsletter recommendations and performance. Section 3 contains the analysis of stock-selection performance. Section 4 analyzes the consistency and relative precision of the three performance-evaluation models. Section 5 looks for short-term persistence, or "hot hands", in newsletter stock-selection ability. Section 6 concludes the paper. Two appendices supplement the text: Appendix A discusses the calculation of newsletter returns and Appendix B describes the construction of the return series used in the DGTW characteristic-matching model.

2. Data

The Hulbert Financial Digest (HFD) has been tracking the performance of investment newsletters since June 1980. It accomplishes this task by subscribing to the print editions and regularly calling the free "telephone hotlines" of the newsletters that have them. The sample includes many well-known newsletters whose editors are often quoted in the financial press or noted for their mutual fund management. Some examples (with editors given in parentheses) are *The Granville Market Letter* (Joe Granville), *Louis Rukeyser's Wall Street* (Louis Rukeyser), *MPT Review* (Louis Navellier), *The Ruff Times* (Howard Ruff), and *The Value Line Investment Survey* (Value Line Publishing, Inc.).

Many newsletters in the HFD sample recommend more than one portfolio; in this case HFD tracks the recommendations for each portfolio separately. Although not all newsletters give specific advice about asset allocation, HFD has adopted a consistent methodology for translating vague advice into "model portfolios" for each newsletter. Details of this methodology are given in Appendix A. The resulting database includes every transaction in these model portfolios. These transactions data are the first useful feature of the HFD sample.

The HFD does not track all newsletters. With a few exceptions, newsletters are added to the database only on January 1 of each year; these annual additions are based on suggestions from HFD subscribers and staff. When first added to the database, no data from prior years is "back-filled": the holdings for each portfolio start on the day that HFD begins tracking it. The database is also free of survivor bias: it includes all portfolios that have ever been tracked by HFD, whether or not the portfolios still exist. Back-fill and survivor biases plague many studies of performance evaluation, and their absence here is the second useful feature of the HFD sample.

Note that although newsletters must survive long enough to be noticed in order to get into the sample, this is not the same thing as "survivor bias". Survivor bias refers to a specific statistical problem induced when data are omitted for historical members of the sample who are not alive at the end of the sample period. This is not the case for the HFD database; data exist for all of the historical members of the sample, whether or not they are alive at the end. Rather, the necessity of getting noticed in order to be tracked by HFD results in a non-random sample – this by itself would not induce bias unless some element of this non-random selection is correlated with subsequent performance and also ignored by the analysis. In this case, there would be a sample-selection bias, but not a survivor bias.[4] To the extent that such a sample-selection bias does exist in the HFD database, the

[4] An example of each kind of bias is the following. Take the universe of all mutual funds that ever existed between 1970 and 1990. Now, the subsample of such funds that survived until 1990 would suffer from survivor bias. Next, consider the subsample of all funds that existed in 1980. Let the analysis of this subsample be based only on their returns *subsequent* to 1980. As long as all of the funds initially in this subsample (i.e., as of 1980) were included, regardless of whether they survived all the way to 1990, then the subsample would not have survivor bias but would possibly suffer from sample-selection bias. This latter bias would occur if survivorship in 1980 had predictive power for subsequent returns, and this relationship was otherwise unexplained by the performance-evaluation model. See Carhart (1997b) for a detailed discussion of these issues. Further problems can occur if funds come in and out of the sample and returns data exist only since their most recent entry; this "re-emerging bias" – which is not a problem for the HFD sample – is discussed in Goetzmann and Jorion (1996).

inclusion of the newsletters that did not get noticed would be likely to make the newsletter performance look even worse than it does in this paper.

The newsletter returns calculated for this paper are different from returns that would be obtained by actually trading on all newsletter recommendations. First, the analysis ignores all issues of allocation across broad asset classes and the timing of these allocations; only the equity portions of the portfolios are tracked and recommendations of all other assets are not included. For this reason, the paper is complementary to Graham and Harvey (1996 and 1997), which focus on the timing decisions between equities and cash. Second, newsletters' short positions are not followed; all portfolios are assumed to be long 100 percent in equities at all times.[5] Third, transactions costs are not considered. For all of these reasons, the returns calculated for this paper do not represent what a real-world investor could expect to achieve by following the advice of the newsletters. The main goal of the paper is to evaluate raw equity performance, and the return calculations are designed for that purpose.

Table I illustrates the growth in HFD's coverage over time.[6] For the entire sample period of June 1980 to December 1996, 153 separate newsletters are covered, with the average newsletter having returns for 81 months and holding 25 stocks in its portfolio. The second column shows the steady increase in HFD's coverage, from 15 newsletters in 1980 to 93 in 1996. The average return for all existing newsletters in any given year is given in the third column. This average return is calculated in several steps. First, I calculate returns for every "model portfolio" recommended by every newsletter. To combine the returns of

[5] Since short positions are excluded, the analysis can only measure the performance of long recommendations; all the results of this paper should be considered with that caveat. Since short positions make up less than 8 percent of all positions, the bulk of the recommendations are still included. Appendix A discusses the reason for this exclusion.

[6] All tables and figures are given at the end of the paper.

different model portfolios within a newsletter, I use an annual rebalancing procedure: for each newsletter, I assume that it invests equally in each of its model portfolios on January 1 of every year; no further rebalancing is done until the following January 1 and each model portfolio is assumed to reinvest all of its returns in its own portfolio. Thus, the weight on each model portfolio will change throughout the year. These steps yield a daily-return series for each newsletter. Finally, to calculate a return on the whole sample of newsletters, I use the same annual rebalancing procedure; each newsletter receives an equal weight on January 1 and reinvests its returns in its own holdings. This yields a return series for the whole sample on every day.[7] Since the annual return on this series is just the average annual return for all active newsletters, I refer to it throughout the paper as the "average" return series. The fourth column contains the value-weighted-market return (VWM) for that year, where VWM is the total return on the CRSP NYSE/AMEX/Nasdaq index. The percentage of newsletters that beat the VWM for that year is given in the fifth column; these results suggest that many newsletters follow a high-beta strategy, since fewer portfolios (usually) beat the market in low VWM years than in high ones.

What kinds of stocks are these newsletters recommending? To answer this question, I look at three different stock characteristics: size (market equity), book-to-market ratio, and momentum (11-month past return, lagged one month). For each characteristic, all stocks with the necessary data are ranked and placed into quintiles. I form quintile breakpoints using only NYSE stocks for size and book-to-market, with these breakpoints used to sort

[7] Further details of these return calculations and their possible biases are discussed in Appendix A.

all CRSP-listed stocks into one of the five ordered categories.[8] The momentum quintiles are formed directly from all CRSP-listed stocks. Then, for each newsletter, I calculate a value-weighted quintile ranking for each characteristic on each day. For example, if on one day a newsletter has half of its weight in stocks that belong to size quintile 5 and half of its weight in stocks that belong to size-quintile 4, the newsletter would receive a size ranking of 4.5 for that day. The time-series average of these daily rankings is the characteristic ranking for the newsletter. By this method, all newsletters receive a 1 to 5 ranking for each characteristic.

Table II summarizes the results for the 153 newsletters. The second column gives the results for the size rankings. Here, a ranking of 1 would indicate that all of a newsletter's recommendations are for stocks in the smallest quintile (by NYSE breakpoints), and a ranking of 5 would indicate that all the recommendations are for stocks in the largest quintile. The table shows that 23 newsletters have average size rankings less than or equal to 2 ("low"), and 41 newsletters have rankings greater than or equal to 4 ("high"). The remaining 89 newsletters have rankings between 2 and 4 ("medium"). By comparison, the S&P 500 has a size ranking of 4.93, a level exceeded by only two newsletters. Thus, even though newsletters are not focused on the smallest stocks, their average stock recommendations are still considerably smaller than this widely followed index.

The third column of Table II shows the results for the book-to-market rankings: 45 newsletters concentrate on low book-to-market "glamour" stocks (average quintile less than

[8] Details of these rankings are given in Appendix C. The use of NYSE breakpoints used to sort *all* stocks into size and book-to-market quintiles, which results in a very different number of stocks across the size "quintiles", is done to maintain consistency with similar sorts done by other authors. See, for example, Fama and French (1993) and Lakonishok, Shleifer and Vishny (1994).

or equal to 2), whereas only 5 newsletters concentrate on high book-to-market "value" stocks (average quintile greater than or equal to 4). Finally, the fourth column shows that 38 newsletters have high momentum rankings, indicating a concentration on the stocks in the highest quintile of past returns, while not a single newsletter focuses on the out-of-favor stocks from the lowest quintile of past returns. The effect of these different strategies on newsletter returns plays an important role in the analysis of the next section.[9]

3. Evaluating Equity Performance: Methodology and Results

In order to properly evaluate stock-selection ability, it is necessary to define a performance-evaluation methodology. Unfortunately, the lack of consensus on the "right" model of expected returns puts performance-evaluation researchers in a quandary: without a generally accepted model of expected returns, it is obviously difficult to define abnormal returns and to quantify the value of investment advice. I deal with this problem by using several different models for expected returns and showing that the evidence on newsletter performance is qualitatively similar in each case.

3.1 The CAPM

The first model of expected returns is the standard CAPM:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i RMRF_t + \epsilon_{i,t} \ , \tag{1}$$

where $R_{i,t}$ is the return on newsletter i in month t, $R_{f,t}$ is the risk-free return in month

[9] It would also be interesting to examine a newsletter's subscription base as a function of returns and/or strategy, but unfortunately, no reliable data on subscriptions are available.

t, and $RMRF_t$ is the month-t value-weighted market return minus the risk-free rate. The estimated intercept, α_i, is the key performance measure from the model. Although the last twenty years have witnessed significant evidence against this unconditional version of the CAPM, it is still used for performance evaluation, both by academics and practitioners.[10] Thus, it provides a good base case for the analysis.

In the first test, I use the average newsletter return as the dependent variable. This return series is the same one as given in the third column of Table I; all existing newsletters receive equal weight on January 1 and then revinvest their returns in themselves throughout the year. The second column of Table III gives the results of estimating (1) for this average return series. The R^2 of the regression is .884. The estimated $\alpha = -.12$ ($= -12$ basis points, per month), with a standard error of .13, so it is not significantly different from zero. The estimated $\beta = 1.20$, reinforcing the evidence of Table I that showed newsletters doing relatively better in bull markets than in bear markets.

I next estimate (1) separately for each newsletter, so that each newsletter has an estimate of α_i and β_i. The results are summarized in the second column of Table IV. Under the assumption that the returns from each newsletter are independent (an assumption that will be dropped later), one can perform tests on the percentiles of the α_i distribution.[11] The idea here is to test whether too many newsletters seem to have extreme good or extreme bad

[10] The "anomalies" literature begun by Basu (1977) and Banz (1981) has now grown quite large. A recent summary can be found in Campbell, Lo, and MacKinlay (1997), chapter 5.7. Some examples of the CAPM's continuing role in performance evaluation can be found in Malkiel (1995), Morningstar (1996), and Shirk et al. (1997). One way to deal with the anomalies is to add more factors or adopt a matching approach based on stock characteristics. These approaches will be considered in sections 3.2 and 3.3, respectively. Another approach, not adopted in this paper, is to develop a conditional version of the CAPM (Jagannathan and Wang (1995)) and use it for performance evaluation (Ferson and Schadt (1996)).

[11] A similar procedure is used in Malkiel (1995) (Table III), Ferson and Schadt (1996) (Table III), and Graham and Harvey (1996) (Table II).

performances. For example, even if returns were generated completely by chance, it would not be surprising if 1 or 2 newsletters out of 153 had α_i estimates that were significantly higher than zero at the 99 percent level of confidence. But what would be surprising? More formally, under the null hypothesis that "true" $\alpha_i = 0$ for all i, then each newsletter has a p_i-value corresponding to its α_i estimate and, for any value of p^*, the number of newsletters with p_i-values below p^* is distributed binomial $(p^*, 153)$. We can then use the binomial distribution to compute the probability of observing the estimated number of newsletters above or below any p^*. Examples of this calculation are given below. Note that the p_i values are derived from one-sided tests of the null hypothesis that $\alpha_i = 0$, and thus $p_i < .5$ and $p_i > .5$ imply that $\alpha_i < 0$ and $\alpha_i > 0$, respectively.

Table IV gives the results for $p^* = .001, .01, .05, .5, .95, .99$, and $.999$. An asterisk (*) next to a table entry denotes that there are too many newsletters that meet the relevant condition, where "too many" is defined as a two-tailed (binomial distribution) probability of less than 5 percent. Thus, an asterisk means that the null hypothesis ($\alpha_i = 0$) can be rejected at the 95 percent level of confidence. Two asterisks (**) imply an analogous rejection at the 99 percent level. Asterisks occur for two cases: $p^* = .01$ and $.05$. Under the null hypothesis that $\alpha_i = 0$ for all i, we would expect 1 percent of newsletters to fall in the $p_i \leq .01$ box; that is, $.01 \times 153 = 1.53$ newsletters. The actual sample has 5. To test whether this is too many, one computes the probability of obtaining 5 or more successes in 153 trials of a binomial distribution with parameter $p^* = .01$. In this case, this probability is .02. One can double this probability (for a two-tailed test) and still reject the null hypothesis ($\alpha_i = 0$ for all i) at a 95 percent level of confidence. For $p_i < .05$ one would expect $.05 \times 153 = 7.515$;

the actual sample has 17. The probability of at least 17 successes with $p^* = .05$ and 153 trials is .002; this implies rejection of the null hypothesis at a 99 percent level of confidence. None of the other boxes in this column of Table IV allow for rejection. One cannot draw firm conclusions from this method, but there is certainly no suggestion of positive excess performance.

Table IV also gives the high and low extreme-values of p_i. Under the null hypothesis, in a sample of N (independent) newsletters, the probability that the highest p_i is at least p is given by $1 - p^N$. Similarly, the probability that the lowest p_i is at most p is given by $1-(1-p)^N$. Table IV also includes these calculations for the sample of $N = 153$ newsletters. The highest p_i is .9881; values at least this high would be expected 84 percent of the time in samples of this size. The lowest p_i is .0015; values at least this low would be expected 21 percent of the time. Neither the highest nor the lowest p_i seem unusual. If anything, the results of Table IV suggest that newsletters underperform relative to the CAPM.

The 153 separate CAPM regressions allow only the testing of percentiles one at a time, and the interpretation of these tests rely on the independence of excess returns across newsletters. For a more complete and formal test of newsletter performance under the CAPM, it is necessary to test all percentiles simultaneously and also to allow for return dependence. In order to calculate a full set of return covariances for such a test, one needs a panel of newsletters with a sufficient number of overlapping months. This requirement forces the exclusion of newsletters that have not survived for a long enough portion of the sample period. This exclusion could induce a positive survivorship bias for the remaining newsletters.[12] If good stock-selection performance aids survival over the necessary sub-period – a relationship established in section 5 – then the surviving newsletters will look better than the whole sample. The implications of this bias are discussed below.

Using time periods of the last 8, 10, 12, 14 and the full sample length of 16.5 years, I form subsamples of newsletters that have survived for every month of the relevant period and estimate (1) in a seemingly-unrelated-regression (SUR) framework. Gibbons, Ross, and Shanken (1989) (GRS) derive an exact finite-sample F-test for the null hypothesis that the alphas of this estimation are jointly equal to zero.[13] The third column of Table V gives the right-tail p-values for the GRS test. Other things equal, the more extreme are the α estimations (both positive and negative), the lower will be the p-value for this test. A p-value of .05 would imply rejection at the 95 percent level of confidence. In these tests, the lowest p-value is .34 for the 14-year subsample (16 newsletters, 168 months), and all of the other subsamples have p-values of at least .79. Clearly, in no case would one reject that the CAPM adequately describes these subsamples of newsletter returns. Since this is a joint test of both the model and the null hypothesis on subsamples that may suffer from survivorship bias, the non-rejection is quite striking. It is possible, of course, that unbiased subsamples would have resulted in rejection due to *underperformance*, and that the biased subsamples have prevented this rejection. In any case, there is no evidence of superior performance.

[12] In this case, the survivorship bias takes the form of a "look-ahead bias" as opposed to a "survivor bias". Look-ahead bias is similar to survivor bias, but is a result of a testing procedure (requiring a minimum return history) rather than a property of the overall sample. Most testing procedures will induce at least some look-ahead bias. See Carhart (1997b) for a detaled discussion of these biases.

[13] The GRS test-statistic is much like an "ordinary" F-statistic for testing joint restrictions, except that it explicitly adjusts for the covariance of the estimates and the reduction in degrees of freedom resulting from the calculation of these covariances. See also Campbell, Lo, and MacKinlay (1997).

3.2 Carhart's 4-factor Model

One line of attack on the unconditional CAPM is that it cannot explain differences in returns for portfolios sorted by stock characteristics such as size, measures of "value" such as the price-to-earnings, cash-flow-to-price, or book-to-market ratios, or past returns (momentum).[14] In light of such evidence, researchers have used many different multifactor models in performance-evaluation studies, and no industry standard has yet emerged. Given the evidence presented in Section 2 indicating the focus of many newsletters on low book-to-market stocks with high past returns, it is important for a study of stock-selection ability to make an attempt to adjust for differences that may result from these strategies. The 4-factor model introduced by Carhart (1997a) attempts to capture the above CAPM anomalies and has proven useful in recent studies of mutual funds (Carhart (1997a), Chevalier and Ellison (1996), and DGTW (1997)). The model is estimated by

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,1} RMRF_t + \beta_{i,2} SMB_t + \beta_{i,3} HML_t + \beta_{i,4} PR1_t + \epsilon_{i,t} \ , \qquad (2)$$

where $R_{i,t}$, $R_{f,t}$, and $RMRF_t$ are defined as in equation (1) and SMB_t, HML_t and $PR1_t$ are the month-t returns to zero-investment factor-mimicking portfolios designed to capture size, book-to-market, and momentum effects, respectively.[15] Equation (2) is a straightforward extension of the Fama-French (1993) 3-factor model, a model which cannot adequately explain the returns of portfolios sorted by past return (Fama and French (1996)). Since the 4-factor model lacks a rigorous theoretical basis, one cannot necessarily give a risk

[14] See Basu (1977) (P/E ratio), Banz (1981) (size), Fama and French (1993) (size and book-to-market), Lakonishok, Shleifer and Vishny (1994) (several value measures), and Jegadeesh and Titman (1993) (momentum).

[15] I am grateful to Mark Carhart for supplying the data on factor returns. For details on how these factors are constructed, see Fama and French (1993) and Carhart (1997).

interpretation to the factors. Thus, I view this regression as a method of performance *attribution* and interpret the estimated alphas as returns in excess of what could have been achieved by passive investment in the "factors". It must be stressed that the debate over the risk interpretation of such multifactor models is ongoing, and the analysis here does not claim to add anything to it.

As with the CAPM, I estimate the 4-factor model for the average newsletter return, for each newsletter independently, and finally in a SUR framework. The third column of Table III shows the results for the average newsletter return. The R^2 of .970 indicates that the 4-factor model mimics these returns very closely. Here, newsletters perform slightly better than under the CAPM, with a positive but statistically insignificant point estimate of α = .03. The negative coefficient on HML is consistent with the evidence of Table II and suggests an overall concentration on low book-to-market stocks. The positive coefficient on SMB may seem puzzling in comparison to the focus on mid-size to large-size stocks that was shown in Table II. The coefficient sign should be expected, however, because even the second-largest quintile of NYSE stocks (quintile 4) actually has a significantly positive loading on SMB.[16] The insignificant coefficient on $PR1$ is somewhat surprising, considering the number of newsletters found to have portfolios weighted towards stocks with high past returns; the (univariate) momentum focus suggested by Table II does not seem evident in the (multivariate) setting of the 4-factor model. Finally, even in this multi-factor setting,

[16] See Fama and French (1996), Table I. The reason for this somewhat paradoxical result is that the "B" part of SMB, which stands for "Small minus Big", is a value-weighted average dominated by the very largest stocks, while "S" is a value-weighted average actually dominated by mid-sized NYSE stocks. (Recall that the breakpoints for these categories are formed using only NYSE stocks.) Thus, mid-sized NYSE stocks will in fact have positive loadings on SMB and only the very largest stocks will have negative loadings. Since the average quintile ranking of the newsletter sample is only slightly above 3 (mid-size), the positive SMB loading makes sense.

the coefficient on $RMRF$, the market "beta", is still significantly above 1.

Next, I estimate (2) for each newsletter independently. Out of the full sample of 153 newsletters, 151 have sufficient return histories for this test.[17] The results are given in the third column of Table IV. Here, newsletter performance looks better than it did under the CAPM. In particular, 15 newsletters have p_i-values greater than .95. This is significantly more than one would expect to observe by chance; in a binomial test with 151 trials and a parameter of .95, the probability of 15 or more "failures" is just over 1 percent. Also, both the $p_i \geq .50$ and $p_i \geq .99$ boxes have more members than would be expected by chance, although neither number is significant at the 95 percent level. The extreme values, however, are not outside of their normal range; the highest p_i, .9959, would be expected to occur about 46 percent of the time in a sample of 151 (independent) newsletters. Nevertheless, this is the first suggestion of any superior performance. The problem, of course, is that the returns are not independent.

The SUR estimation attempts to control for the problem of return dependence. As in section 3.1, subsamples are formed by taking all newsletters that have survived for the last 8, 10, 12, 14 and 16.5 years, and then (2) is estimated in a SUR framework. The p-values for the GRS F-statistics are given in the fourth column of Table V. Despite the potential survivorship bias in these subsamples, the null hypothesis that the alphas are jointly equal to zero is not rejected for any of the subperiods. The closest the test comes to rejection is for the 1983 - 1996 subperiod, with a right-tailed p-value of .37. Thus, once one adjusts

[17] To perform inference on the 5 coefficients of the model, one needs at least 6 months of data. This data requirement can induce survivorship bias for the remaining newsletters. (See Carhart (1997b)). Since only 2 newsletters are dropped and neither had extreme performance under the CAPM, this bias is likely to be small or nonexistent.

for return dependencies and tests the whole distribution of alphas, there is no significant evidence of superior performance.[18]

3.3 Daniel, Grinblatt, Titman and Wermers' (DGTW) Characteristic-Matching Model

In addition to the return anomalies, there are other difficulties with interpreting the alphas from factor-model regressions. As has been pointed out by several authors (Grinblatt and Titman (1995), Ferson and Schadt (1996), DGTW (1997)), estimated alphas and betas will be biased if factor loadings are correlated with factor realizations, with the direction of these biases dependent on the sign of the correlation. If, for example, a newsletter were to consistently increase its holdings of small stocks (and, thus, its loading on SMB) just before periods of high returns for SMB, then this positive correlation would serve to bias upwards its coefficient on SMB, with the effect on its estimated α depending on the sign of the average return to SMB over the sample period.

To solve this problem, one could attempt to correctly measure the factor loadings at all times and make adjustments accordingly. If the only available data are portfolio returns, then such a measurement is not possible. With the HFD sample, however, each transaction is known. Therefore, it is possible, in principle, to find "matching" portfolios at all times. In the example described above, such a matching portfolio would increase its holding of small stocks at precisely the same time as the underlying portfolio, and biases due to market

[18] An earlier version of this paper (Metrick (1997)) used the newsletters' "model portfolios" as the main unit of study, rather than the approach taken by this paper of collapsing all of a newsletter's portfolios down to a single return. Although performance looked better with portfolios as the unit, many of the best performers were drawn from a high number of nearly identical portfolios of 2 newsletters. There, SUR tests were not possible because the number of portfolios was too large to estimate return covariances. Since interpretation of those results was problematic, the unit of analysis was shifted to newsletter averages.

timing might be reduced or avoided.[19] A similar approach is often adopted for event studies, with stock-by-stock matching based on characteristics such as size, industry, or the book-to-market ratio.[20] DGTW (1997) extend this approach to a performance-evaluation study and use characteristic matching on every stock held by mutual funds in quarterly portfolio "snapshots". I adopt a slightly modified version of the DGTW characteristic-matching model to obtain the paper's third performance measure.

To obtain the DGTW measure, one begins by constructing 125 "bins" through a 5x5x5 sort on size, book-to-market, and momentum quintiles. As in the calculations for Table II in section 2, NYSE breakpoints are used for size and book-to-market, and NYSE/AMEX/Nasdaq breakpoints are used for momentum, with *all* NYSE/AMEX/Nasdaq stocks placed into quintiles on the basis of these breakpoints. These three characteristics correspond to the SMB, HML, and $PR1$ factors, respectively. Next, one calculates a daily return for each bin; within each bin, stocks are equally-weighted on the first day of each month and are assumed to revinvest their returns in their own stock throughout the month. Details on these sorts and return calculations are given in Appendix B. In this model, each stock falls into a bin, which I call its "matching bin". Abnormal returns are then measured as the difference between a stock's return and its matching-bin's return.[21]

For newsletters, the monthly measure of abnormal returns is calculated as the return

[19] Another attempt to solve this problem is to adjust for changes in expected factor realizations rather than factor loadings. Ferson and Schadt (1996) take this approach and employ conditional asset-pricing models that use publicly available predictors of factor realizations.

[20] See, for example, Brav and Gompers (1997), Desai and Jain (1993), Ikenberry, Lakonishok and Vermaelen (1995) and Womack (1996) for recent examples using these characteristics.

[21] The choice of these characteristics for bin formation is an attempt to maximize the explanatory power for expected returns while also keeping the number of bins to a manageable number. Nevertheless, it must be acknowledged that the use of these 125 bins is arbitrary, and is used here mainly for consistency with the original model of DGTW (1997).

on a zero-investment portfolio that is long in the newsletter's actual portfolio and short in a portfolio constructed using equivalent weights in the matching bins. In effect, one just combines the monthly abnormal returns for each stock in the portfolio. Not all of a newsletter's stocks will be included in this calculation; if a stock cannot be matched to a bin, then it will not be included in the test. The two main reasons for a failure to match are, first, insufficient past returns for a momentum calculation and, second, the absence of a book-equity observation in Compustat. Both of these data requirements lead to new issues being deleted from the portfolios. To the extent that new issues underperform similar stocks, this will cause an upward bias in the estimated selectivity performance measure.[22] If such an upward bias exists, it does not seem to have a significant effect on the results. For example, when I repeat the CAPM and 4-factor tests of Table III using returns calculated only from stocks that have bin assignments, then the results are almost identical: CAPM $\alpha = -14$ basis points (2 points lower than $\alpha = -12$ in Table III) and 4-factor $\alpha = 5$ basis points (2 points higher than $\alpha = 3$ in Table III).

The assumption underlying this model is that all stocks in the same bin have exactly the same expected return. If this assumption is satisfied, then the performance measure will have a zero expected return *at all times.* Thus, newsletters that shift their portfolio composition conditional on expected factor realizations will have no bias in their estimated performance measure.

To see how newsletter abnormal returns are calculated, consider the following example.

[22] See Loughran and Ritter (1995) and Brav and Gompers (1997) for evidence on the new-issues bias. On a "bin-adjusted" basis, Brav and Gompers (1997) work suggests that this bias should not be large. See also Chan, Jegadeesh, and Lakonishok (1995) for a discussion of the bias induced by omitting stocks that do not have data in Compustat.

Newsletter XYZ recommends holdings of 40 percent in IBM and 60 percent in Exxon on March 1 and does not recommend any further transactions in March. Suppose that this newsletter earns a return of 3 percent in March. During this time, IBM belongs to bin 122 and Exxon belongs to bin 124. Then, form another portfolio on March 1 that holds 40 percent of its holdings in bin 122's stocks (equally-weighted) and 60 percent of its holdings in bin 124's stocks (equally-weighted). Suppose that this bin portfolio earns a return of 2 percent for March. In this case, the abnormal return for XYZ in March would be $3 - 2 = 1$ percent. If Newsletter XYZ had instead shifted its portfolio during the month, then the bin allocations would have been shifted at the same time.

To represent this formally, some notation is needed:

$d \in t$: the set of all days d in month t;

$s \in i$: the set of all stocks s held by newsletter i;

$b(s)$: bin b matched to stock s;

$R_{s,d}$ = net return on stock s on day d;

$R_{b(s),d}$ = net return on bin b matched to stock s on day d;

$W_{s(i),d}$ = weight placed on stock s by newsletter i on day d;

$R_{i,t}$ = net return to newsletter i in month t;

$R_{b(i),t}$ = net return to the bins b matched to the stocks in newsletter i in month t;

Then, for each month, the net return on newsletter i is given by

$$R_{i,t} = \prod_{d \in t} \left(1 + \sum_{s \in i} (W_{s(i),d} * R_{s,d}) \right) - 1 \qquad (3)$$

where $\sum_{s \in i} (W_{s(i),d} * R_{s,d})$ is the actual net return for newsletter i on day d. The net return to the bins b matched to the stocks in newsletter i in month t is given by

$$R_{b(i),t} = \prod_{d \in t} \left(1 + \sum_{s \in i} (W_{s(i),d} * R_{b(s),d}) \right) - 1 \tag{4}$$

where $\sum_{s \in i}(W_{s(i),d} * R_{b(s),d})$ is the net return that would be achieved on day d if all funds were invested in the bins matched to the recommended stocks of newsletter i. Now, define $CS_{i,t}$, the overall characteristic-selectivity measure for newsletter i in month t, as

$$CS_{i,t} = R_{i,t} - R_{b(i),t} \tag{5}$$

The analogue to the factor models' α can be found by estimating a regression of $CS_{i,t}$ on an intercept term. The intercept from this regression, analogous to the α_i in the factor models, is called CS_i. There are two other equivalent ways to think about the CS_i calculation, each of which can aid intuition. First, instead of a regression of $CS_{i,t}$ on an intercept term, one can estimate a regression of $R_{i,t}$ on an intercept and $R_{b(i),t}$, where the coefficient on $R_{b(i),t}$ is constrained to be equal to 1. This regression allows for a sensible interpretation of R^2 as the fraction of the return variance explained by the characteristic-matching model, and thus is the estimation used below for the average-return series. Second, one can interpret $CS_{i,t}$ as the monthly return on a trading strategy that is always long in the underlying newsletter stocks and short in the matching bins. Then, CS_i is just the average monthly return to this trading strategy, and can be calculated as

$$CS_i = \overline{R}_i - \overline{R}_{b(i)} \tag{6}$$

where $\overline{R}_i$ is the average monthly return to the newsletter and $\overline{R}_{b(i)}$ is the average monthly return to its corresponding bins. All of these methods compute exactly the same estimates for CS_i.

As with the factor models, I use the characteristic-matching model on the average newsletter return, on each newsletter independently, and in a SUR framework using all newsletters that have been in existence for various subperiods. The results for the average return are presented in the fourth column of Table III. For purposes of computing an R^2 for this regression, I use the regression of $R_{i,t}$ on an intercept term and $R_{b(i),t}$, where the coefficient on the latter is constrained to be equal to 1. The standard error of the intercept term, CS, is .06, which is lower than the equivalent standard errors on the α estimates in the CAPM and 4-factor model (.13 and .07, respectively). This means that one *would* have been able to reject the null of no excess performance if CS were found to have been greater than 12 basis points in absolute value. The actual estimate of $CS = .01$ is not significantly different from zero.

I next analyze the individual CS_i estimates for each newsletter. The results are given in the fifth column of Table IV. The results appear "normal". Under the assumption that returns are independent, none of the tails of this distribution seem unusually heavy. The worst performer has a t-statistic of -3.50 for 192 months of data. This implies a p_i-value of .0003; one would expect to see a p_i at least this low in only about 4 percent of samples of this size. The highest p_i is .9920; one would expect to see a p_i at least this high in about 71 percent of samples. None of these results suggest superior performance.

I next drop the assumption of independence for the subsamples of 8, 10, 12, 14, and 16.5

year (whole sample) survivors. The GRS p-values for the SUR estimations on these samples are given in the fourth column of Table V. As with the CAPM and the 4-factor model, one finds that the null hypothesis of all alphas being jointly equal to zero cannot be rejected for any of the subperiods, with no p-value less than .59. These non-rejections occur despite the potential survivorship bias in the subsamples. Overall, under the characteristic-matching model there is no significant evidence of superior performance.

It could be argued that "real" stock-picking ability is demonstrated not just by earning excess returns relative to bins, but also through successful timing *across* bins. For example, a newsletter may not be a "timer" in the traditional sense, but it may act on a belief that small stocks are undervalued relative to large stocks. If this belief turns out to be correct, then this skill would likely be missed by the CS statistic. DGTW define a measure called "characteristic timing" to capture such effects. I omit the formal definition and analysis of this measure, because in unreported tests I find no evidence of any superior performance. If anything, newsletters exhibit below-average characteristic timing.

4. A Comparison of Performance-Evaluation Models

4.1 Consistency Across Models

The results presented above do not support positive abnormal performance for any of the three models employed. In this subsection, I examine the consistency of the results across models. The Venn diagrams in Figure 1 illustrate the results for the extreme performers in each model. In Panel A, each newsletter with at least one extreme "good" performance is included. A good performance is defined as inclusion in at least one of the $p_i \geq .95$ boxes in

Table IV: a left-tail probability of at least 95 percent for a one-sided t-test that α_i (for either the CAPM or 4-factor model) or CS_i (for the characteristic-matching model) are greater than 0.

We can read the entries in Figure 1 to see that only a single newsletter had extreme good performances under all 3 models, while 6 newsletters had extreme good performances under both the 4-factor model and characteristic-matching models (but not the CAPM), and 2 newsletters met the criterion for both the CAPM and 4-factor model (but not for the characteristic-matching model). By examining the reasons for disparate performance across the models. we can gain insight into the role that model choice plays in performance evaluation. The famous *Value Line Investment Survey* is an interesting case study. For the 168 months that this newsletter is in the sample (1/83 - 12/96), it achieved an annualized return of 21.9 percent.[23] The CAPM α_i is 22 basis points per month, with a standard error for this estimate of 19 basis points. Thus, the t-statistic is not high enough to be significant by the $p_i > .95$ criterion. Under the 4-factor model, the *Value Line* newsletter portfolio shows a positive loading on SMB, a negative loading on HML, and a positive loading on $PR1$. As compared to the CAPM, these loadings slightly improve the measured α_i to 32 basis points per month, and the improved fit of the regression (Adjusted R^2 of .91 vs. .84 for the CAPM) causes the standard error on this estimate to fall to 15 basis points. The resulting t-statistic for a one-sided test on $\alpha_i > 0$ is high enough to land *Value Line* into the $p_i > .95$ box in Table IV. Here, it is both the effect on the point estimate of α_i *and* its improved precision which causes *Value Line*'s performance to look better under the 4-

[23] *Value Line* is an example of the "no-backfilling" rule of the HFD sample. *Value Line* has been providing recommendations on this portfolio since 1965, but only the years since 1983 (when HFD began contemperanous coverage) are included in my study.

factor model than under the CAPM. For the characteristic-matching model, the estimated CS_i is 21 basis points per month, but the standard error of this estimate is 11 basis points per month. This added precision allows *Value Line* to demonstrate significantly positive performance with its CS_i measure.

Panel B of Figure 1 shows the analogous categories of extreme "bad" performers under the three models. A bad performance is defined as inclusion in one of the $p_i \leq .05$ boxes in Table IV. Again, it is useful to look at a case where the measures might lead to different conclusions. The *RHM Survey of Warrants* has significantly negative abnormal performance under the CAPM, with an estimated α_i of -99 basis points per month (t-statistic = -2.26) over the 156 months that it is in the sample. As it turns out, the data suggest that this newsletter focused on relatively low-momentum small growth stocks – a terrible strategy over its lifespan. When its return history is analyzed using the 4-factor model, it is found to load positively on SMB, negatively on HML, and negatively on $PR1$ – all three of these loadings were bad for realized returns over the relevant period. Once adjusted for these loadings, the newsletter has a 4-factor α_i of -60 basis points per month (t-statistic = -1.37); this is still low, but not low enough to land in the $p_i < .05$ category. The results for the CS_i measure are a bit better, with an estimate of -28 basis points per month (t-statistic = -0.75). Of course, *RHM* is still responsible for its poor (or unlucky) choice of strategy, but the analysis can remain agnostic as to whether this strategy earned lower returns for risk or non-risk reasons. Nevertheless, the use of several different models does allow one to separate out the "equity style" from the "stock-selection" components of *RHM*'s performance.

4.2 Comparing Precision Across Models

A model comparison on a case-by-case basis is helpful for understanding individual newsletter performance, but it cannot say anything general about the relative precision across models. There are certainly specific cases where precision seems to increase as one moves from the CAPM to the 4-factor model to the characteristic-matching model, but is this true in general? Should a researcher who has transactions data bother with a characteristic-matching model, or would a factor model be just as powerful? With the advent of transactions-based databases in performance evaluation, these questions have important practical implications.[24] It is possible to answer these questions using simulated portfolios, but this may miss important elements of actual managed portfolios. For example, managed portfolios often follow focused size, value, momentum, or industry-based strategies that may evolve over time. If simulated portfolios do not reflect these strategies, then the results will not accurately capture the relationship between models; recall that the characteristic-matching model is designed specifically to capture such effects. It seems useful to analyze the precision of these models "in the field"; that is the purpose of this section.

The goal of this analysis is to construct measures of precision than can be compared across models. To derive these measures, consider the test statistics for any given newsletter under each model. For the CAPM,

$$t_{CAPM_i} = \frac{\alpha_{CAPM_i}}{\sigma_{CAPM_i}} \tag{7}$$

[24] For recent examples of transactions databases in performance evaluation analyses, see Eckbo and Smith (1998), and Jeng (1998) for insider trading, Barber and Odean (1998) and Grinblatt and Keloharju (1998) for individual investors, and DGTW (1997) for mutual funds.

where σ_{CAPM_i} is the standard error of the α_{CAPM_i} estimate. For the 4-factor model.

$$t_{4F_i} = \frac{\alpha_{4F_i}}{\sigma_{4F_i}} \tag{8}$$

where σ_{4F_i} is the standard error of the α_{4F_i} estimate. For the characteristic-matching model,

$$t_{CS_i} = \frac{CS_i}{\sigma_{CS_i}} \tag{9}$$

where σ_{CS_i} is the standard error for the characteristic-selectivity measure, CS_i. To motivate the measure of precision, consider an experiment in which each stock for a given newsletter will have a small and equal increase for every month it is held by the newsletter. This induces an equivalent increase in the monthly newsletter return, $\overline{R}_i$. What would be the effect of this experiment on the test statistic for each model? In other words, what would the derivative of the test statistics be with respect to $\overline{R}_i$?

Since the factor returns are calculated from a large number of stocks, one can assume that the increase in the returns of a single newsletter's stocks would have only a negligible effect on the factor returns. Therefore, for both factor models, an increase in an average newsletter return, $\overline{R_i}$, induces a one-for-one increase in α_i, so the effect on the test statistics are

$$\frac{\partial t_{CAPM_i}}{\partial \overline{R}_i} = \frac{1}{\sigma_{CAPM_i}} \tag{10}$$

for the CAPM, and

$$\frac{\partial t_{4F_i}}{\partial \overline{R}_i} = \frac{1}{\sigma_{4F_i}} \tag{11}$$

for the 4-factor model. These measures of precision are simply the inverses of the standard errors for the respective α_i estimates. For the characteristic-matching model, the procedure is slightly more complicated. First, (6) can be substituted into (9) to yield

$$t_{CS_i} = \frac{\overline{R}_i - \overline{R}_{b(i)}}{\sigma_{CS_i}}, \tag{12}$$

where $\overline{R}_{b(i)}$ is the mean monthly return of the bins matched to the stocks in newsletter i. To compute the derivative of this t-statistic with respect to $\overline{R}_i$, we need to consider separately each term in the numerator. The derivative of $\overline{R}_i$ with respect to itself is obviously 1, but what of the second term? A logical criticism of DGTW's characteristic-matching approach is that the use of "too many" bins will effectively render CS_i to be meaningless. In the extreme, of course, this criticism must be true. If one were to choose as many bins as there are stocks, then by definition all $\overline{R}_i$ would be equal to $\overline{R}_{b(i)}$ and all CS_i would be zero. But is 125 bins "too many"? If the bins do not contain many stocks, or if a newsletter owns most of the stocks in any one bin, then the effect of our experiment on the bin returns could be substantial. Consider the effect of the experiment for a "representative" first day of a month – this eliminates the need for time subscripts and simplifies the weighting of stocks within bins.[25] Let $n_{s,b}$ be the number of stocks in the bin b that contains stock s. If stock s has an increase in its return, then its effect on the corresponding bin's return, $\overline{R}_{b(i)}$, will be proportional to $\frac{1}{n_{s,b}}$. Therefore, I write λ_i, the derivative of $\overline{R}_{b(i)}$ with respect to $\overline{R}_i$, as

$$\lambda_i = \frac{\partial \overline{R}_{b(i)}}{\partial \overline{R}_i} = \sum_{b=1}^{125} \sum_{s \in i} \left(W_{s(i)} * \sum_{s \in i} \frac{1}{n_{s,b}} \right). \tag{13}$$

[25] After the first of the month, weighting of stocks within bins will not be equal. Thus, the analysis below is only an approximation on those days. Please see Appendix B for a discussion of bin construction.

where $W_{s(i)}$ is the weight placed on stock s by newsletter i. In an extreme case where there is only one stock in every bin (i.e. all $n_{s,b} = 1$ and $\overline{R}_i = \overline{R}_{b(i)}$), then λ_i will be equal to 1; if there are very few bins (i.e. large $n_{s,b}$), then λ_i will be close to zero. With (13) in hand, one can differentiate (12) with respect to $\overline{R}_i$; this yields the total effect on the t-statistic for CS_i of an increase in the return for every stock in newsletter i:

$$\frac{\partial t_{CS_i}}{\partial \overline{R}_i} = \frac{1 - \lambda_i}{\sigma_{CS_i}} \tag{14}$$

(10), (11), and (14) are measures of precision for each test. I next compare these measures. The ratio of (11) to (10) leads to a natural measure of relative precision between the 4-factor model and the CAPM:

$$\text{relative precision (4-factor / CAPM)} = \frac{\frac{\partial t_{4F_i}}{\partial \overline{R}_i}}{\frac{\partial t_{CAPM_i}}{\partial \overline{R}_i}} = \frac{\sigma_{CAPM_i}}{\sigma_{4F_i}} \tag{15}$$

For the 151 newsletters for which both models can be estimated, 81 have a relative precision greater than 1; in these cases, the 4-factor model provides a more precise estimate than the CAPM. The median measure in this sample is just below 1.01. This relative precision is higher for newsletters with long return histories or many stocks in their portfolios. In a regression of relative precision (or its logarithm) on length of return history and "average number of stocks held", the coefficients on both independent variables are positive and significant at the 95 percent level. As an illustration of this effect, if one considers only those newsletters with at least 10 years of data, then the median measure of relative precision is 1.03, and 27 out of 40 newsletters have a measure greater than 1. Similarly, among newsletters that hold an average of 20 or more stocks in their portfolios, the median measure is 1.11, with 43 out of 61 measures greater than 1. Clearly, the benefits of the 4-factor model over the CAPM do not become apparent unless newsletters have long histories and/or hold many stocks.[26]

I next turn to a comparison between the 4-factor and characteristic-matching models. First, it is necessary to redefine the set of stocks so that they are the same for both models. In section 3.2, the 4-factor regressions are estimated on newsletter returns using all stocks, while the estimates of CS_i in section 3.3 are (necessarily) done only on the subset of stocks that have bin-assignments. To make an apples-to-apples comparison, one must re-estimate the 4-factor model on the same returns as used for the characteristic-matching model. The results are qualitatively similar to those presented in section 3.2. Dividing (14) by (11) yields a relative precision for the two models of

$$\text{relative precision (Characteristic-Matching / 4-factor)} = \frac{\frac{\partial t_{CS_i}}{\partial \overline{R_i}}}{\frac{\partial t_{4F_i}}{\partial \overline{R_i}}} = \frac{(1-\lambda_i)\sigma_{4F_i}}{\sigma_{CS_i}} \qquad (16)$$

An estimate of relative precision can be computed for 150 newsletters.[27] Estimates of λ_i are computed on July 1 of every year, and then averaged across all years to form a single estimate for each newsletter. For the whole sample, the median λ_i is .069. The median measure of relative precision for the 150 newsletters is 1.10, with 126 of the newsletters having

[26] This analysis ignores differences in the degrees of freedom of each test statistic. Since the 4-factor model estimates 3 more parameters than the CAPM, this will make the 4-factor model relatively less precise than the measure in (15) suggests. For newsletters with more than 30 months of data, these effects would usually be negligible.

[27] 151 newsletters have adequate histories when all stocks are used to compute returns (section 3.2), but 1 newsletter drops to only 4 months of returns when the bin-assignments are required.

ratios greater than 1.[28] Overall, relative precision in (16) is not related to length of return history nor to the average number of stocks held. For example, among newsletters with at least 10 years of history, there is a median measure of 1.09 and 34 out of 40 newsletters have measures greater than 1. Furthermore, there are no significant coefficients for a regression of relative precision (or its logarithm) on length of return history and average number of stocks held.

There are two main conclusions from this exercise. First, the gain in precision for the 4-factor model over the CAPM is small or nonexistent for newsletters with few stocks and short return histories, but grows with history length and number of stocks. Second, a characteristic-matching approach offers the potential for significant gains in precision over the 4-factor-model. These gains are across the board: they are not a function of return history or average number of stocks held by newsletters. This finding is worthy of further study on other real and simulated data, with extensions to a wider class of factor and characteristic-based methods.

5. Do Newsletters Get "Hot Hands"?

The tests of section 3 are designed to assess the stock-selection of newsletters over their entire lifetime. The results show no significant evidence of superior performance. However, what if some newsletters can get "hot hands" and make successful stock selections for short periods? Suppose that instead of picking a newsletter and sticking with it for its entire existence, one were to choose only the best performers over the previous year – could a strategy

[28] As in (15), the measure in (16) does not make degree of freedom corrections. Such corrections would tend to make the comparison more favorable to the characteristic-matching model, which estimates only 1 parameter vs. 5 in the 4-factor model. These effects would usually be neglibible for newsletters with more than 30 months of history.

like this earn excess returns for stock selection? Specifically, consider a trading strategy that divides the entire newsletter universe into deciles based on their raw returns from the previous year. Then, each January, reset the portfolio to include only the newsletters from the highest decile. The annualized return to this strategy would have been 15.7 percent for the 1981 - 1996 period, which is slightly higher than the 14.4 percent earned by the CRSP value-weighted index. Conversely, the annualized return to the strategy that invested in the lowest decile would have been only 8.9 percent over the same time period. The success of this simple short-term persistence strategy would appear to lend support to the existence of hot hands. But is this a real persistence of superior performance, or can it be explained by factor loadings?

A similar hot-hands effect is well documented in other contexts. Graham and Harvey (1996) find that the market-timing recommendations of newsletters whose previous prediction was "correct" are significantly better than those of newsletters whose previous prediction was "incorrect", but they also find that this persistence is not strong enough to identify any long-term overperformance. Brown and Goetzmann (1995) demonstrate a pattern of short-run persistence in mutual funds, concentrated among poor performers, and provide evidence that this persistence is due to common strategies that are not captured by standard models. Hendricks, Patel and Zeckhauser (1993) show that a strategy of buying the best past performers and selling the worst performers earned significant positive returns from 1974 to 1988 within a sample of no-load mutual funds. Carhart (1997a) confirms this result but shows that almost all of the difference can be explained by the 4-factor model (equation 2), with the only remaining persistence occurring for the worst-performing funds.

In this section, I adopt Carhart's methodology and analyze the returns of decile-sorted past performers.

Before turning to the factor models, it is helpful to review newsletter survivorship and the persistence of the decile rankings themselves. Figure 2 shows a histogram of newsletter "death" frequencies in year t as a function of the year $t-1$ raw-return decile. The overall trend suggests that good performance helps survival. The top half (deciles 1 through 5) have an average death rate of 5 percent, and the bottom half (deciles 6 through 10) have an average death rate of 10 percent. A logit estimation of survival (1 if yes, 0 if no) on the previous year's decile yields a point estimate of .15 with a standard error of .05. This estimate can be translated into probability terms, where it implies a 1 percent survival advantage in year t for each higher decile in year $t-1$.[29] These results are consistent with Graham and Harvey's (1996) finding that good market-timing performance by newsletters increases their survival probabilities. Also, in samples of mutual funds. the impact of performance on survivorship is large and significant.[30] Despite the fact that my performance measure is focused purely on stock selection, ignores transactions costs, and is not itself a published measure available to newsletter or HFD subscribers, there is still a correlation with survival.

Similarly, the decile rankings also show some persistence. The correlation of year $t-1$ deciles with year t deciles is .09, a positive relationship that is significant at the 1-percent level. This result is consistent with the relative success of the trading strategy of buying the top performing decile in each year. However, one still doesn't know if the persistence is

[29] The point estimate from the logit regression can be (approximately) converted to probability terms by multiplying it by $p*(1-p)$, where $p = .93$, the sample mean survival rate (see Maddala (1983)). This yields an answer of $.15 * .93 * .07 = .0098 \approx 1.0$ percent.

[30] For evidence on this point, see Brown and Goetzmann (1995), Malkiel (1995) and Carhart (1997b).

explained by "betas" (factor loadings) or "alphas" (stock-picking skill).

Table VI summarizes the results of 4-factor regressions (equation 2) on each of the hot-hands deciles, and also on a zero-investment portfolio long in the highest-performing decile and short in the lowest-performing decile.[31] The $PR1$ coefficients match well with intuition, with significant positive coefficients for deciles 1 and 3, and significant negative coefficients for deciles 9 and 10. The last row of Table VI shows the regression estimates when decile 1 minus decile 10 returns are used as the dependent variable. The estimated α is -16 basis points per month. Under the 4-factor model, decile 10 appears to outperform decile 1. What happened to the excess performance reported at the top of this section? A large part of it is attributable to the momentum factor. The average return to $PR1$ over this sample period is 85 basis points *per month*. A loading of .55 on $PR1$ attributes 47 basis points per month by itself. In addition, decile 1 has a higher loading on $RMRF$ and HML than does decile 10; both these factors had positive returns over the sample period. Of all the α estimates, only the α for decile 7 is significant at the 95 percent level. In a SUR framework, the p-value for the GRS F-statistic for the 10 deciles is .51, well above any critical value. Although there is mild short-term persistence in raw performance, there is no evidence that newsletters exhibit short-term persistence of *abnormal* performance.

6. Conclusion

As greater numbers of individuals participate in equity markets, it becomes increasingly important for financial economists to form scientific opinions about the behavior and per-

[31] The tests in this section are all carried out using rankings on past raw returns rather than rankings based on past performance measures. This is done in part to stay consistent with the literature, but also because past performance measures, if biased, will then bias the persistence tests as well. Please see Carhart (1997a) for a discussion of this point.

formance of retail investors and the advisors they follow. In such instances, our theories and intuition are often strong, but the data availability is weak and empirical results are rare. This paper has attempted to partially fill this gap by studying the rich and carefully constructed HFD database of 153 investment newsletters. In addition, the unique properties of the database allow for insights into the methodology of performance evaluation; these insights suggest significant benefits for using transactions data as a supplement to returns data.

Several different methodologies are employed to evaluate stock-picking skill by newsletters: a pair of returns-based factor models, the CAPM and the 4-factor model of Carhart (1997a), and a transactions-based approach, the characteristic-matching model of Daniel, Grinblatt, Titman and Wermers (1997). Each model is estimated on an average newsletter return, on each newsletter independently, and on dependency-adjusted subsamples of long-surviving newsletters. Overall, newsletters do not demonstrate significant abnormal performance: average abnormal returns are close to zero, the best performing newsletter under each model does not seem unusual given the sample size, and, for the CAPM and characteristic-matching model, there are not "too many" good performing newsletters. Only under the 4-factor model does there appear to be too many good performers, and even this weak result disappears once return dependencies across newsletters are explicitly considered.

In addition to these tests on the whole history of newsletter returns, I also test for short-term persistence of returns: do newsletters get "hot hands"? To test for persistence, all newsletters are placed into decile portfolios based on their previous year's return. When the 4-factor model is estimated on 16 years of returns to these decile-sorted portfolios, the estimated set of alphas is insignificantly different from zero. Although a strategy of buying the best past performers and selling the worst ones would have earned positive raw returns over the sample period, the abnormal returns would have been negative and insignificantly different from zero. Overall, there is no evidence of superior stock-selection skill by newsletters, either over short or long horizons.

The main methodological contribution of the paper takes the form of a "field-study"; in this sample of newsletters, I measure the relative precision of the three models: CAPM, 4-factor, and characteristic-matching. I find that the latter approach, which requires transactions data, shows a median improvement in precision of 10 percent over the 4-factor model. This compares favorably with a median improvement of 1 percent for the 4-factor model over the CAPM. The 4-factor model shows its greatest improvement over the CAPM for newsletters with long histories and many stocks in their portfolios; even for this group the characteristic-matching model is more precise. These results suggest that researchers with transactions data should attempt to exploit this property when they do performance evaluation.

7. References

[1] Basu S., 1977,"The Investment Performance of Common Stocks in Relation to Their Price-to-Earnings: A Test of the Efficient Markets Hypothesis", *Journal of Finance* **32**: 663-682.

[2] Banz R., 1981, "The Relation Between Return and Market Value of Stocks", *Journal of Financial Economics*, **38**: 269-296.

[3] Barber, Brad M. and Douglas Loeffler, 1993, "The 'Dartboard' Column: Second-Hand Information and Price Pressure", *Journal of Financial and Quantitative Analysis* **28**, 273-284.

[4] Barber, Brad M. and Terrance Odean, 1998, "The Common Stock Investment Performance of Individual Investors", Manuscript, Graduate School of Management, UC Davis, May.

[5] Black, Fischer, 1973, "Yes Virginia, There is Hope: Tests of the Value Line Ranking System", *Financial Analysts Journal*, **29**: 10-14, September.

[6] Brav, Alon and Paul Gompers, 1997, "Myth or Reality: The Long-Run Underperformance

of Initial Public Offerings: Evidence from Venture and Nonventure Capital-backed Companies", *Journal of Finance* **52**(5): 1791-1822.

[7] Brimelow, Peter, 1986, *The Wall Street Gurus*. Random House: New York.

[8] Brown, Stephen, and William N. Goetzmann, 1995, "Performance Persistence", *Journal of Finance* **50**: 679-698, June.

[9] Brown, Keith C., W.V. Harlow, and Laura T. Starks, 1996, "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry", *Journal of Finance* **51**: 85-109, March.

[10] Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press: Princeton, NJ.

[11] Canina, Linda, Roni Michaely, Richard Thaler, and Kent Womack, 1998, "Caveat Compounder: A Warning about Using the Daily CRSP Equal-Weighted Index to Compute Long-Run Excess Returns", *Journal of Finance* **53**: 403-416, February.

[12] Carhart, Mark, 1997a, "On Persistence in Mutual Fund Performance", *Journal of Finance* **52**: 57-82, March.

[13] Carhart, Mark, 1997b, "Mutual Fund Survivorship", Mimeo, Marshall School of Business, University of Southern California, May.

[14] Chan, Louis K.C., 1995, Narasimhan Jegadeesh and Josef Lakonishok, "Evaluating the Performance of Value Versus Glamour Stocks: The Impact of Selection Bias", *Journal of Financial Economics*, **38**: 269-296.

[15] Chevalier, Judith and Glenn Ellison, 1996, "Are Some Mutual Fund Managers Better than Others? Cross-Sectional Patterns in Behavior and Performance", NBER Working Paper #5852, December.

[16] Copeland, T. and D. Mayers, 1982, "The Value Line Enigma (1965-1978): A Case Study of Performance Evaluation Issues", *Journal of Financial Economics*, **10**: 289-321, November.

[17] Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, "Measuring Mutual Fund Performance with Characteristic Based Benchmarks", *Journal of Finance* **52**: 1035-1058, July.

[18] Daniel, Kent and Sheridan Titman, 1997, "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns", *Journal of Finance* **52**: 1-34, March.

[19] Desai, Hemang, and Prem C. Jain, 1995, "An Analysis of the Recommendations of the 'Superstar' Money Managers at Barron's Annual Roundtable", *Journal of Finance* **50**: 1257-1274, September.

[20] Eckbo, B. Espen and David C. Smith, "The Conditional Performance of Insider Trades", *Journal of Finance* **53**: 467-498, April.

[21] Fama, Eugene and Kenneth French, 1993, "Common Risk Factors in the Returns on Stocks and Bonds", *Journal of Financial Economics* **33**: 3-56.

[22] Fama, Eugene and Kenneth French, 1996, "Multifactor Explanations of Asset Pricing Anomalies", *Journal of Finance* **51**: 55-84, March.

[23] Ferson, Wayne and Rudi Schadt, 1996, "Measuring Fund Strategy and Performance in Changing Economic Conditions", *Journal of Finance* **51**: 425-461, June.

[24] Gibbons, Michael R., Stephen A. Ross, and Jay Shanken, 1989, "A Test of the Efficiency of a Given Portfolio", *Econometrica* **57**, 1121-1152, September.

[25] Graham, John. R., 1998, "Herding Among Investment Newsletters: Theory and Evidence", Forthcoming in *Journal of Finance.*

[26] Graham, John R. and Campbell R. Harvey, 1996, "Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendations", *Journal of Financial Economics* **42** (3), November.

[27] Graham, John R. and Campbell R. Harvey, 1997, "Grading the Performance of Market Timing Newsletters", Forthcoming in the *Financial Analysts Journal.*

[28] Grinblatt, Mark and Sheridan Titman, 1995, "Performance Evaluation", in *Handbook in Operations Research and Management Science: Volume 9, Finance*, R.A. Jarrow, V. Maksimovic and W.T. Ziemba, editors: Elsevier, Amsterdam.

[29] Grinblatt, Mark and Matti Keloharju, 1998, "Momentum Investing and Performance Using Finland's Unique Data Set", Manuscript, UCLA.

[30] Goetzmann, William N. and Phillipe Jorion, 1997, "Re-emerging Markets", NBER Working Paper #5906.

[31] Hendricks, Darryl, Jayendu Patel and Richard Zeckhauser, 1993, "Hot Hands in Mutual Funds: The Persistence of Performance, 1974 - 1988", *Journal of Finance* **48**: 93-130.

[32] Huberman, G. and S. Kandel, 1990, "Market Efficiency and Value Line's Record", *Journal of Business* **63**: 187-216, June.

[33] Hulbert, Mark, 1996, "Mail-Order Portfolios", *Forbes*, February 26.

[34] Ikenberry, David, Josef Lakonishok, and Theo Vermaelen, 1995, "Market Underreaction to Open Market Share Repurchases", *Journal of Financial Economics* **39**(2&3), November.

[35] Jeng, Leslie A., 1998, "Corporate Insiders, Market Makers, and the Window of Opportunity", Manuscript, Harvard University, May.

[36] Lakonishok, Josef, Andrei Shleifer, and Robert Vishny, 1994, "Contrarian Investment, Extrapolation, and Risk", *Journal of Finance* **49**: 1541-1578.

[37] Lewis, Craig M., Richard J. Rogalski, and James K. Seward, 1997, "The Information Content of Value Line Convertible Bond Rankings", *Journal of Portfolio Management* **24**(1): 42-52, Fall.

[38] Loughran, Tim and Jay Ritter, 1995, "The New Issues Puzzle", *Journal of Finance* **50**: 23 - 52.

[39] Malkiel, Burton, 1995, "Returns from Investing in Mutual Funds 1971 to 1991", *Journal of Finance* **50**: 549-572, June.

[40] MacKinlay, A. Craig, 1995, "Multifactor Explanations do not Explain Deviations from the CAPM", *Journal of Financial Economics* **38**: 3-28.

[41] Maddala, G.S., *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press: Cambridge.

[42] Metrick, Andrew, 1997, "Performance Evaluation with Transactions Data: The Case of Investment Newsletters", Manuscript, Harvard University, July.

[43] Morningstar, 1996, *Ascent* database, April.

[44] Shelton, J., 1967, "The Value-Line Contest: A Test of the Predictability of Stock-Price Changes", *Journal of Business* **40**: 251 - 269, July.

[45] Shirk, George S. III, Julie Cuenca and Greg Carlson, 1997, "Platinum Fund$", *Mutual Funds*, October.

[46] Shumway, Tyler, 1997, "The Delisting Bias in CRSP Data", *Journal of Finance* **52**: 327-340, March.

[47] Shumway, Tyler and Vincent A. Warther, 1997, "The Delisting Bias in CRSP's Nasdaq Data and its Implications for Interpretation of the Size Effect", Manuscript, University of Michigan, June.

[48] Stickel, Scott E., 1985, "The Effect of Value Line Investment Survey Rank Changes on Common Stock Prices", *Journal of Financial Economics* **14**, 121-143.

[49] Womack, Kent, 1996, "Do Brokerage Analysts' Recommendations Have Investment Value?", *Journal of Finance* **51**: 137-168, March.

Appendix A - The Calculation of Newsletter Returns

Newsletters make recommendations in many different forms, and the HFD has devised a methodology to convert advice into "model portfolios". Depending on the year, between one-half and two-thirds of the newsletters give explicit advice about percentages or shares to be held of each security. The remaining newsletters do not give explicit advice, but use some system of ranking stocks by desirability, such as a numerical ranking or a "buy", "hold", "sell" classification. HFD takes such rankings and forms equally-weighted portfolios from the stocks in the most highly recommended category, treating removals from this category as sells. The resulting model portfolios, some exact and some constructed, are the raw material for the paper. At this point, the HFD database leaves off and my analysis begins.

The first thing done with these model portfolios is to purge them of all non-equities and short-sales. Model portfolios are rebalanced on "transaction days" so that they include only the long equity positions. The idea here is to focus on pure equity positions and be able to abstract from market timing (equities vs. other assets). Short positions are ignored because there is no obvious way to include them without allowing implicit market-timing bets. For example, consider a newsletter that recommends a portfolio that is 100 shares (long) in IBM and 100 shares (short) in Microsoft. If Microsoft is more expensive per share than IBM, then this will be a net short position in equities. If the Microsoft position is taken off, then the portfolio becomes net long in equities. The returns to the newsletter will then reflect this market-timing strategy. While the CS measure (section 3.3) would not be affected by this, the factor models (sections 3.1 and 3.2) would be affected; this would complicate the comparison across models. The eliminated short recommendations comprise approximately 7.6 percent of all newsletter positions, and more than half of the newsletters never make a short recommendation. Therefore, the remaining analysis still includes the vast majority of equity recommendations.

To calculate returns to these portfolios, I assume that all trades take place at the first daily closing price after the recommendation was received. This methodology implies that a newsletter with a large market impact would have its returns underestimated. Unfortunately, it is not possible to do a direct test for this bias, because some stocks are recommended through telephone hotlines after major moves have already occurred on that day; any inclusion of "day zero" returns would be confounded by this effect. Since the methodology employed in this paper uses closing prices for the day that recommendations were received in the mail (or by phone before the market close), it is at least in principle an "implementable" strategy, albeit one without transaction costs. Only those stocks contained in the CRSP daily files are included in the return calculation. Because most stocks do not have delisting returns in CRSP, I substitute the estimated delisting returns suggested by Shumway

(1997) and Shumway and Warther (1997): -30 percent for performance-related delists on the NYSE/AMEX and -55 percent on the NASDAQ. At the end of this procedure, each model portfolio has a daily return series for its whole lifetime.

One possible source of bias in portfolio returns are the differences in the ways stocks are recommended. Newsletters that recommend exact share amounts for every stock should have well-constructed returns. However, newsletters that recommend only portfolio weights present some problems. For example, consider a newsletter which recommends a holding of 50 percent in IBM and changes the other 50 percent of the portfolio every day. Then, this portfolio will effectively gain the benefits of bid-ask bounce in IBM, with a daily rebalancing of its portfolio weight back to 50 percent. In the extreme, daily rebalancing can cause large differences in measured returns (see Canina, et al. (1998)). To see if this bias is evident in the sample, I compare the t-statistics in each model for "percent-type" vs. "share-type" newsletters. To keep t-statistics comparable, I restrict this comparison to newsletters that have at least two years of data. The results are that, for each model, the share-type newsletters have *higher* mean t-statistics than the percent-type newsletters, although these differences are never statistically significant. Thus, it seems unlikely that rebalancing bias is playing a large role. This is probably because very few percent-type newsletters transact often enough to induce a measurable effect.

Since many newsletters have multiple portfolios, the next step is to create "newsletter returns" from these portfolio returns. This step is necessary because several newsletters have multiple portfolios with nearly identical holdings, and this duplication leads to difficulties in the interpretation of results on the full sample of portfolios. The results of this paper are quantitatively similar and qualitatively identical for several different methods of combining portfolio returns into newsletter returns. In the method used in the text, newsletter returns are calculated as an annually-rebalanced average of their portfolio returns. That is, each newsletter is assumed to invest equally in all its portfolios on January 1 of each year. All portfolios revinvest their own proceeds, so that weight on each portfolio shifts throughout the year. This yields a daily return series for each newsletter. The resulting annual newsletter return is the same as the average annual return across all of its portfolios.

Appendix B - Formation of the Quintiles and the 125 Bin Returns

Each July 1, all stocks are sorted into size and book-to-market quintiles using NYSE breakpoints. Size is computed at June month-end, and the book-to-market ratio is calculated as a stock's most recent fiscal year's book equity divided by the previous December-end market equity. Stocks that lack any of the necessary measures in CRSP (for market equity) or Compustat (for book equity) are excluded. REITs, ADRs, units of beneficial trust, and negative book-equity stocks are also excluded. The use of NYSE breakpoints causes the number of stocks in the smallest size quintile, and to some extent the second smallest quintile, to be greater than those in the other quintiles. The momentum quintiles are recalculated monthly based on prior 11-month return lagged one month. This is the same definition of momentum as in Carhart (1997a) and DGTW (1997); as pointed out in the latter paper, the omission of the most recent month is done to avoid problems of bid-ask bounce and monthly return reversals (Jegadeesh (1990)). Momentum quintile-breakpoints are formed using *all* stocks. All stocks tied on a quintile breakpoint are assigned to the lower of the two quintiles. A "bin" is the intersection of the three independent quintile sorts. Size and book-to-market quintiles do not change from July to June, but stocks can still change bins once a month when their momentum quintiles change. The returns to each bin are then computed from a monthly (*not* daily) equal-weighting procedure. Specifically, the weights for each stock are set to be equal on the first day of the month, and then a stock's total return is assumed to be revinvested in its own shares through the month. Therefore, each stock's weight is always proportional to its "month-to-date" return. This procedure is motivated by a suggestion in Panina et al. (1998), and provides the benefits of equal weighting (no small stock effect within bins) while minimizing the compounding bias indices if daily equal-weighted returns are used. Where appropriate, I use the same estimated delisting returns as described in Appendix A. The bin returns are used as the $R_{b(s),d}$ in equation (3) and Section 3.3. The resulting calculation of CS_i includes only stocks that have the data necessary to place them into bins. The omission of stocks that lack the necessary data should not be a source of bias: See Chan, Jegadeesh and Lakonishok (1995), Brav and Gompers (1997), and the CAPM and 4-factor regression results discussed in section 3.3.

Table I: Summary Statistics on Newsletters

Total newsletters in sample: 153
Average newsletter coverage (months): 81
Average # of stocks in newsletter portfolios: 25

Year	Newsletters Followed	Average Newsletter Return	VWM Return	% Newsletters Beating VWM Return
1980	15	32.57	27.25	80.0
1981	21	-5.60	-3.95	52.4
1982	24	29.03	20.42	70.8
1983	34	21.40	22.70	47.1
1984	44	-6.24	3.28	22.7
1985	57	32.72	31.47	52.6
1986	59	13.51	15.60	47.5
1987	65	0.28	1.81	35.4
1988	69	16.18	17.61	44.9
1989	69	26.45	28.45	49.3
1990	76	-9.43	-6.02	42.1
1991	79	47.77	33.59	67.1
1992	81	11.32	9.03	54.3
1993	88	15.75	11.48	60.2
1994	88	-3.23	-0.62	39.8
1995	90	36.34	35.73	47.8
1996	93	21.20	21.24	45.1

Table I gives summary statistics for the HFD sample. Column 2 shows the total number of newsletters that existed for at least part of the year. Column 3 shows the annual return for an average of all newsletters. This series is rebalanced annually and includes all existing newsletters at all times. Weights are normalized to one for each newsletter on January 1 of each year, with newsletter returns reinvested in their own portfolios. Thus, at all times the weight on each newsletter will be equal to its "year-to-date" return. Column 4 shows the value-weighted-market return for the year: the total return on the CRSP index (NYSE\AMEX\Nasdaq). (All returns for 1980 are calculated from June 1, when the HFD data begins, to December 31.) Column 5 shows the fraction of newsletters that beat the value-weighted-market return. If a newsletter was only in the market for part of the year, then its return is only compared to the value-weighted-market return for that part.

Table II – Newsletters Sorted by the Average Size, Book-to-Market, and Momentum Quintiles of their Portfolios

	Size	B/M Ratio	Momentum
Low (rank ≤ 2)	23	45	0
Medium (2< rank <4)	89	103	115
High (rank ≥ 4)	41	5	38

Table II summarizes the size, book-to-market ratio and momentum characteristics of the stocks recommended by the 153 newsletters. Each July, all stocks with available data are placed in size and book-to-market quintiles, with quintile breakpoints made using only NYSE stocks. Momentum quintiles are formed monthly using past 11-month returns lagged one month; all stocks in CRSP are used to form momentum breakpoints. (Please consult Appendix B for a more complete description of how these rankings are made.) Average quintiles for each characteristic and newsletter are calculated daily by multiplying each stock's portfolio weight (in a newsletter portfolio) by its (numerical) quintile, and summing across all stocks in the portfolio. Column 2 separates newsletters into low (small size, average quintile less than or equal to 2), medium (average quintile between 2 and 4), and high (large size, average quintile greater than 4). Column 3 performs a similar classification for the book-to-market quintile, and column 4 does so for momentum.

Table III – Tests on Average Returns from all Newsletters

	CAPM	4-factor Model	Characteristic-Matching Model
α	-0.12 (0.13)	0.03 (0.07)	
RMRF	1.20** (0.03)	1.09** (0.02)	
SMB		0.60** (0.03)	
HML		-0.19** (0.03)	
PR1		-0.02 (0.03)	
CS			0.01 (0.06)
Adjusted R^2	0.884	0.970	0.973

Table III presents the results of tests on the average of all newsletter returns. The series is rebalanced annually and includes all existing newsletters at all times. Weights are normalized to one for each newsletter on January 1 of each year, with newsletter returns reinvested in their own portfolios. Thus, at all times the weight on each newsletter will be equal to its "year-to-date" return. Column 2 gives the results for the CAPM (equation 1); column 3 gives the results for the 4-factor model (equation 2); column 4 gives the results for the characteristic-matching model (equations 3-6). α is the regression intercept, and the next four rows give coefficients and standard errors (in parentheses) for the independent variables: RMRF, SMB, HML, and PR1. These variables are the returns to zero-investment portfolios designed to capture market, size, book-to-market, and momentum effects, respectively. (Please consult Fama and French (1993) and Carhart (1997a) on the construction of these factors.) CS is the Characteristic-Selectivity measure, calculated here as the intercept of a regression of $R_{i,t}$ (equation 3) on an intercept term and $R_{b(i),t}$ (equation 4), where the coefficient on the latter is constrained to be 1. (**) indicates two-tail significance at the 99 percent level.

Table IV – Tests on All Newsletters Treated Independently

	CAPM	4-factor Model	Characteristic-Matching Model
Number of newsletters = N	153	151	153
lowest p_i	0.0015	0.0076	0.0003
$1-(1-p_i)^N$ (lowest p_i)	0.21	0.69	0.04
$p_i \leq .001$	0	0	1
$p_i \leq .01$	5*	1	2
$p_i \leq .05$	17**	11	9
$p_i \geq .50$	66	78	82
$p_i \geq .95$	3	15*	11
$p_i \geq .99$	0	4	1
$p_i \geq .999$	0	0	0
highest p_i	0.9881	0.9959	0.9920
$1-p_i^N$ (highest p_i)	0.84	0.46	0.71

Table IV summarizes the results of several tests on the whole sample of individual newsletters. Column 2 gives the results for the CAPM (equation 1); column 3 gives the results for the 4-factor model (equation 2); column 4 gives the results for the characteristic-matching model (equations 3-6). The p_i-values are for one-sided t-tests on $\alpha_i > 0$, (in columns 2 and 3) and on $CS_i > 0$ (in column 4). The entries in rows 4-10 are the number of newsletters with p-values (p_i) that satisfy the condition given in the first column. For each entry in rows 4-10, tail probabilities are computed from the binomial distribution (p, N). (*) and (**) indicate binomial (two-tail) probabilities of less than 5 percent and 1 percent, respectively. Row 3 gives the probability of observing a lowest p-value no higher than p_i (in row 2) in a sample of size N (from row 1). Similarly, row 12 gives the probability of observing a highest p-value of at least p_i (in row 11) in a sample of size N.

Table V - Gibbons, Ross and Shanken (1989) p-values

Subperiod	Number of Newsletters	CAPM	4-factor Model	Characteristic-Matching Model
1989-1996	36	0.92	0.82	0.98
1987-1996	29	0.94	0.51	0.99
1985-1996	25	0.79	0.45	0.85
1983-1996	16[†]	0.34	0.37	0.88
1980-1996	6	0.83	0.66	0.59

Table V summarizes the results from seemingly-unrelated-regressions (SUR) on the returns for a subset of newsletters. The table entries are right-tail p-values for the Gibbons, Ross and Shanken (1989) F-test that all the intercept terms are jointly equal to zero. In each test, only the newsletters that have survived for the entire subperiod are included. Column 2 gives the number of newsletters that survived for the corresponding subperiod listed in column 1. Column 3 gives the results for the CAPM (equation 1); column 4 gives the results for the 4-factor model (equation 2); column 5 gives the results for the characteristic-matching model (equations 3-6). For column 5, the SUR is estimated as the CS measures on a constant.

[†] One of these 16 newsletters does not hold any stocks with bin assignments during some months in 1983; thus, there are only 15 newsletters with a complete set of returns for the characteristic-matching model.

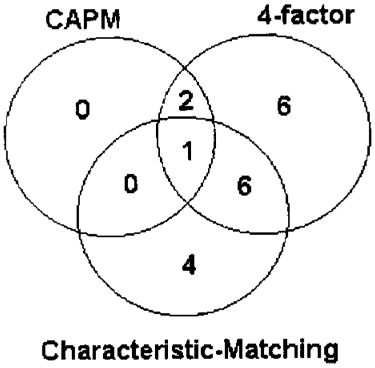
Table VI – 4-Factor Regressions on “Hot-Hands” Deciles

Decile	R^2	α	RMRF	SMB	HML	PR1
1	.81	0.09	1.14**	0.97**	-0.40**	0.26**
2	.71	0.32	1.04**	0.68**	-0.26*	0.08
3	.86	0.04	0.99**	0.63**	-0.22**	0.13*
4	.89	-0.01	1.15**	0.46**	-0.19**	-0.00
5	.83	-0.04	1.02**	0.42**	-0.05	-0.01
6	.88	-0.19	1.09**	0.88**	0.08	0.04
7	.83	-0.42*	1.11**	0.59**	-0.12	0.02
8	.84	0.15	1.13**	0.68**	-0.28**	-0.11
9	.81	0.15	1.15**	0.67**	-0.19*	-0.15*
10	.60	0.25	0.96**	0.45**	-0.61**	-0.29*
1-10	.08	-0.16	0.18	0.52**	0.22	0.55**

Table VI summarizes the results of 4-factor regressions (equation 2) on decile portfolios of newsletters sorted by past return. Each January 1, all newsletters are ranked and placed into deciles based on their previous year’s raw return. Decile 1 has the highest past performance. Decile returns are calculated as the average of their constituent newsletters' returns, with weights on each day equal to a newsletter's “year-to-date” returns. The resulting 10 return series, plus an additional series consisting of decile 1 minus decile 10, are then regressed on the 4 factors of equation 2. Column 2 gives the adjusted R^2 for each regression. In column 3, α is the regression intercept, and the next four columns give coefficients for the independent variables: RMRF, SMB, HML, and PR1. These variables are the returns to zero-investment portfolios designed to capture market, size, book-to-market, and momentum effects, respectively. (Please consult Fama and French (1993) and Carhart (1997a) on the construction of these factors.) (*) and (**) indicate two-tail significance at the 95 percent level and 99 percent level, respectively.

Figure 1: Model Comparison for Extreme Performers

Panel A. Extreme "Good" Performers: p ≥ .95



Panel B. Extreme "Bad" Performers: p ≤ .05

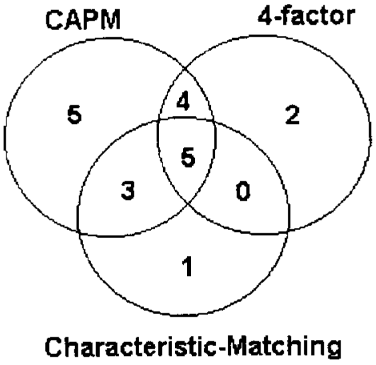


Figure 1 illustrates the number of newsletters with extreme performance across the three performance-evaluation models. The Venn diagram in Panel A shows the extreme "good" performers: those newsletters with a left tail probability greater than or equal to 95 percent for $\alpha_i > 0$ (in the CAPM and 4-factor model) or $CS_i > 0$ (in the characteristic-matching model.) This cutoff corresponds to the eighth row of Table IV ($p_i \geq .95$), and the total in each circle in Panel A equals the entry for the corresponding column in the eighth row of Table IV. Panel B analogously shows the extreme "bad" performers: ($p_i \leq .05$ on any of the tests.) This cutoff corresponds to the sixth row in Table IV.

Figure 2 – Newsletter Death Frequencies

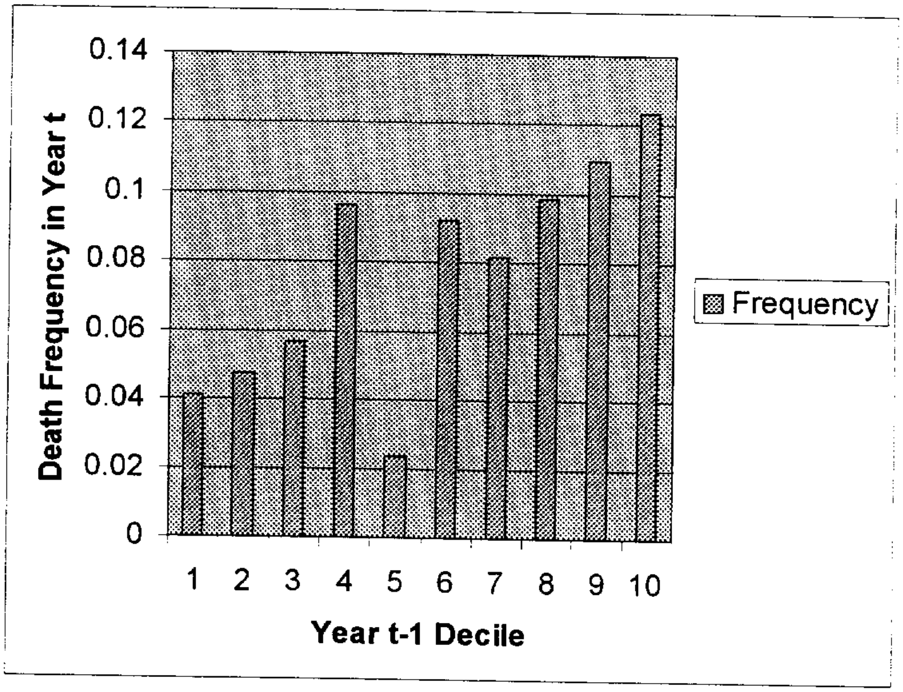


Figure 2 plots newsletter death frequencies by year t as a function of year t-1 decile. At the beginning of each year t, all newsletters are placed into decile rankings by their raw returns in year t-1. If a newsletter ceases publishing at any time before the end of year t, then this is counted as a "death". For the entire sample, the death frequency for year t for each year t-1 decile is computed and plotted in the figure.