# Grading the Performance of Market-Timing Newsletters

John R. Graham and Campbell R. Harvey

*Many investment newsletters offer market-timing advice; that is, they are supposed to recommend increased stock market weights before market appreciations and decreased weights before market declines. Examination of the performance of 326 newsletter asset-allocation strategies for the 1983–95 period shows that as a group, newsletters do not appear to possess any special information about the future direction of the market. Nevertheless, investment newsletters that are on a hot streak (have correctly anticipated the direction of the market in previous recommendations) may provide valuable information about future returns.*

If investment newsletters can "time the market," they should recommend that their subscribers increase the portion of funds invested in the stock market prior to market increases and decrease the portion dedicated to stocks prior to market declines. Graham and Harvey (1994, 1996) found strong evidence that as a group, newsletters cannot time the stock market, but their emphasis was on evaluating newsletters as a group; they did not investigate whether individual newsletters can give valuable investment advice. This study focuses on techniques that can be used to identify superior individual newsletters. Importantly, these techniques can be used to evaluate the performance of a wide variety of investments and are not specific to newsletters.

We studied the recommendations made by 326 investment newsletters between 1983 and 1995, a much larger sample than we used in Graham and Harvey (1996). We first investigated whether, on average, newsletters increase their recommended equity weights prior to market rises and decrease their recommended weights prior to market declines. Our evidence indicates that newsletters, on average, do not alter their recommendations appropriately, but we also identified an intriguing phenomenon: Newsletters that are on a "hot streak" (e.g., they increased equity weights in their previous three recommendations and the stock market yielded a positive return each time) showed substantial ability to time the market, and those that are on a "cold streak" continued to give poor investment advice. This finding suggests that investors can potentially earn superior returns by following "hot" advice.

We investigated whether this apparent "hot hands phenomenon" is misleading in that it implies some newsletters give valuable investment advice, when in fact, they do not provide any new information. For example, if a newsletter makes its recommendations based solely on the index of leading economic indicators, which is available to the public at no cost, then an investor could simply invest based on the leading indicators and avoid paying a newsletter subscription fee. Using a regression analysis that determines whether newsletters provide any information beyond that which is publicly available, we found that between 8 percent and 15 percent of the investment newsletters do provide useful investment advice. The problem is how to identify those superior advisors.

We propose two new performance measures that, after controlling for risk, identify investment advisors that give superior advice.[1] The idea of controlling for risk is very important but sometimes overlooked. For example, an advisor could recommend that an investor margin his or her investment so as to be 200 percent exposed to the market at all times. Such an investment would yield approximately twice the return from going 100 percent the market, but it would also subject an investor to substantial risk. Our performance measures adjust for risk and consequently penalize a 200 percent long strategy before comparing it with a 100 percent long strategy. The advice given by investment advisors, however, is much more interesting than simply "always go 200 percent long the market," and our measures are designed to evaluate a wide variety of scenarios involving changing asset allocation weights, selecting specific stocks or

*John R. Graham is an assistant professor of finance, and Campbell R. Harvey is the J. Paul Sticht Professor of International Business. Both are at the Fuqua School of Business at Duke University.*

different investment vehicles, and so forth.

We used these new performance measures to identify the superior investment newsletters. We ranked the newsletters by "grading" them (e.g., A+ for the best newsletters, A for the next best, and so forth). We also compared the out-of-sample performance of our measures with a ranking based on the benchmark used in the industry, the Sharpe ratio. Our results suggest that our new grading scheme yields valuable information about future performance. For example, newsletters that achieved an A for the 1986–90 period should (for the 1991–95 period) outperform those that received a B in 1986 to 1990; those that received a B from 1986 to 1990 should outperform those that received a C, and so forth. The results indicate that our measures yield more information about future performance than does the Sharpe ratio.

## DATA

Mark Hulbert of the *Hulbert Financial Digest* has collected data on a large sample of investment newsletters each year since the early 1980s. We used the Hulbert data to form portfolios based on the newsletters' recommendations.

### Asset-Allocation Newsletters

The *Hulbert Financial Digest* provided the data on newsletter-recommended asset allocations from 1983 through 1995. There are 132 newsletters that make recommendations, and many of them offer multiple strategies. Hence, the database covers a total of 326 newsletter strategies. A well-defined recommendation is a proposed portfolio composition that satisfies the property that recommended long equity plus short equity plus cash minus margin is 100 percent of the investment. In almost all cases, the nonequity category is cash, although in some cases it may be fixed income. To simplify the analysis, we assumed that the nonequity investment is represented by the 30-day Treasury bill and that the noncash investment is in the S&P 500 Index. We note, however, newsletters that do not allocate exclusively into equity and cash.[2]

Observations are added to Hulbert's database in three ways. First, the newsletter's recommendation is entered on the day it is received in the mail. Second, if the newsletter has a free hotline, Hulbert calls this number each day to supplement the recommendations received by mail. Third, if the letter has previously expressed a stop-loss position (e.g., sell if the Dow Jones Industrial Average reaches 9000), Hulbert implements this order as a recommendation if the condition occurs.

In contrast to data on mutual funds, our data have essentially no survivorship bias. New funds are added on the day Hulbert first receives the letter and no data are deleted when a newsletter ceases to exist.

The newsletters, in aggregate, provided 31,038 total recommendations, which slightly overstates the number of observations because, before 1993, Hulbert added a year-end and year-beginning forecast for every newsletter in existence, even if its forecast did not change. That is, if the newsletter recommended an 80 percent equity/20 percent cash mix on November 30, 1991, and changed to 70 percent/30 percent in February 1992, Hulbert adds the recommendation of 80 percent/20 percent on December 31, 1991. These additions are innocuous and do not affect any of our results.[3]

In the raw data, an observation can occur on any day during a month, and multiple observations may occur in any month. Our tests, however, are based on monthly recommendations. This approach allows us to link our work to the growing literature on conditional performance measurement, which uses monthly data. To this end, we used the last observation in a month as our monthly asset weight recommendation.

We also added observations for months in which a newsletter was in existence but did not change its forecast. This type of addition is the same as Hulbert makes at the turn of the year. For example, if a newsletter provides only a January forecast in a particular year, we assigned 11 additional monthly observations. These additions have no effect on newsletter performance. If recommendations are made quarterly, the portfolio weights are assumed to be constant over the three months of the quarter. We made one exception to the addition rule: If a letter explicitly withdraws a previous forecast without making a new forecast, we do not carry forward the old forecast. The net result of the deletion of intramonth recommendations and the addition of recommendations is 20,080 observations.

### Implementing Newsletter Recommendations

To track the performance of the newsletters in making recommendations, we formed portfolios consisting of S&P 500 futures and cash. For example, a 50 percent cash/50 percent equity recommendation would be implemented by fully investing the initial principal in a 30-day Treasury bill and taking a long position in the S&P 500 futures equal to 50 percent of the initial principal. The returns consist of the gain on the T-bill plus the price change of the futures contract, all divided by the initial principal. Because the futures market

fined our analysis to the 1983–95 sample. Given that most of the recommendations are implemented in the futures market, our analysis ignores transaction costs.

## PERFORMANCE EVALUATION

There are widely different approaches to the problem of evaluating portfolio performance. We present some new measures that are simple to implement and are rich in economic intuition. We also compare our new measures with the traditional ones.

### Long-Term Performance

The predictive ability of newsletters can be measured in a variety of ways, each providing a different perspective on performance.

*Mean–variance analysis.* Figure 1 shows the long-run performance of the newsletters' recommended portfolios. Each point on the graph represents the average annual return to an S&P 500 futures and T-bill strategy that uses the newsletters' asset-allocation weights. Also depicted on the graph are the average returns to a 100 percent investment in the S&P 500 futures, as well as a 100 percent investment in the 30-day T-bill. A curve connecting these two additional portfolios represents the combinations of equity and cash held in fixed proportion from January 1983 to December 1995 (upper panel) and January 1991 to December 1995 (lower panel). This curve is the efficient frontier. An investor can obtain a return along the efficient frontier without any ability to time the market. A newsletter that can time the market is represented by a point lying above the efficient frontier.

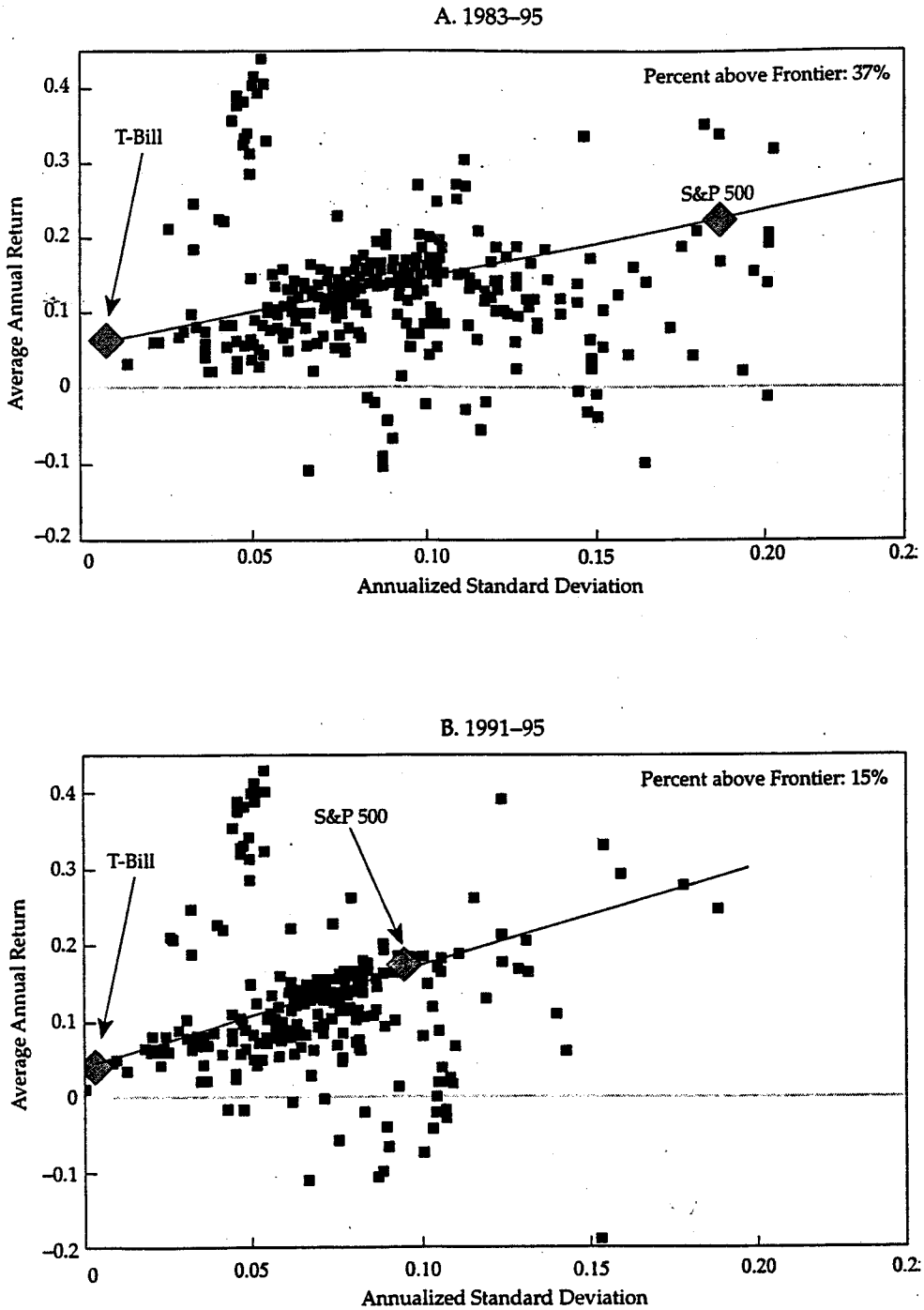
In the whole 1983–95 period, 37 percent of the newsletter portfolios lie above the curve representing constant-weight "passive" strategies (i.e., 37 percent lie above the efficient frontier).[4] This analysis may be misleading, however. Only 10 newsletters existed for the whole sample period; of those, 8 were below the frontier. Thus, in many cases, we are plotting the return and volatility of a newsletter that might have existed for only 1 year against a mean–variance frontier calculated over the past 13 years. For example, three of the newsletters in the All-Star Fund family had an average return of approximately 40 percent and less than 10 percent volatility, but those newsletters existed for only 12 months of the sample—in 1995 (a period when the return on the S&P 500 futures index plus cash was 41 percent and the realized volatility was only 5 percent).

The lower panel of Figure 1 shows the same analysis for the most recent five years of the sample period. Only 15 percent of the newsletter strategies lie above the frontier. In this subsample, the efficient frontier is calculated with five years of data, so the problem of the mean–variance frontier representing a different sample from the period over which a newsletter's performance is calculated is attenuated (relative to the upper panel). It does not eliminate all of the problems, however, because approximately one-third of the newsletters plotted have a sample shorter than five years.

*New performance measures.* To compare each newsletter with an efficient-frontier portfolio calculated over the newsletter's time horizon, we used two new performance measures. Both are designed to compare newsletter performance with a benchmark return—adjusted for risk. To implement the new measures, we first calculated the average annualized returns and volatility of a hypothetical investment fund that follows each newsletter's recommendations for the complete history of the newsletter.[5]

For Measure 1 (GH1), we levered or unlevered the S&P 500 futures to have the exact same volatility as the newsletter (or fund) for the evaluation period, 1991 to 1995. GH1 is the difference between the newsletter return and the return on the volatility-matched futures portfolio. The upper panel of Figure 2 details the results of unlevering the S&P 500 futures index by combining it with the T-bill to match the volatility of Fund A. This strategy produces a much higher return than Fund A has. Hence, GH1 for Fund A is negative, indicating underperformance. Fund B achieves greater performance than a levered S&P futures position and receives a positive GH1. The intuition is simple. If the investor had a target level of volatility equal to that of Fund A, then the investor would have been much better off holding a fixed-weight combination of S&P 500 futures and T-bills than acting on the newsletter's recommendation and (potentially) rebalancing every month.

Measure 2 (GH2) is related to but different from GH1. In this measure, we lever up or down the newsletter's recommended investment strategy (using a T-bill) so that the strategy has exactly the same volatility as the S&P 500. The lower panel of Figure 2 shows the geometry of this measure. If Fund A is levered up to achieve the same volatility as the S&P 500 for the evaluation period, it has a lower average return than a simple "buy-and-hold the S&P 500" strategy. Hence, the GH2 measure is negative. In contrast, if we lever Fund B downward (by combining the newsletter strategy with a cash investment) to achieve the same volatility as the

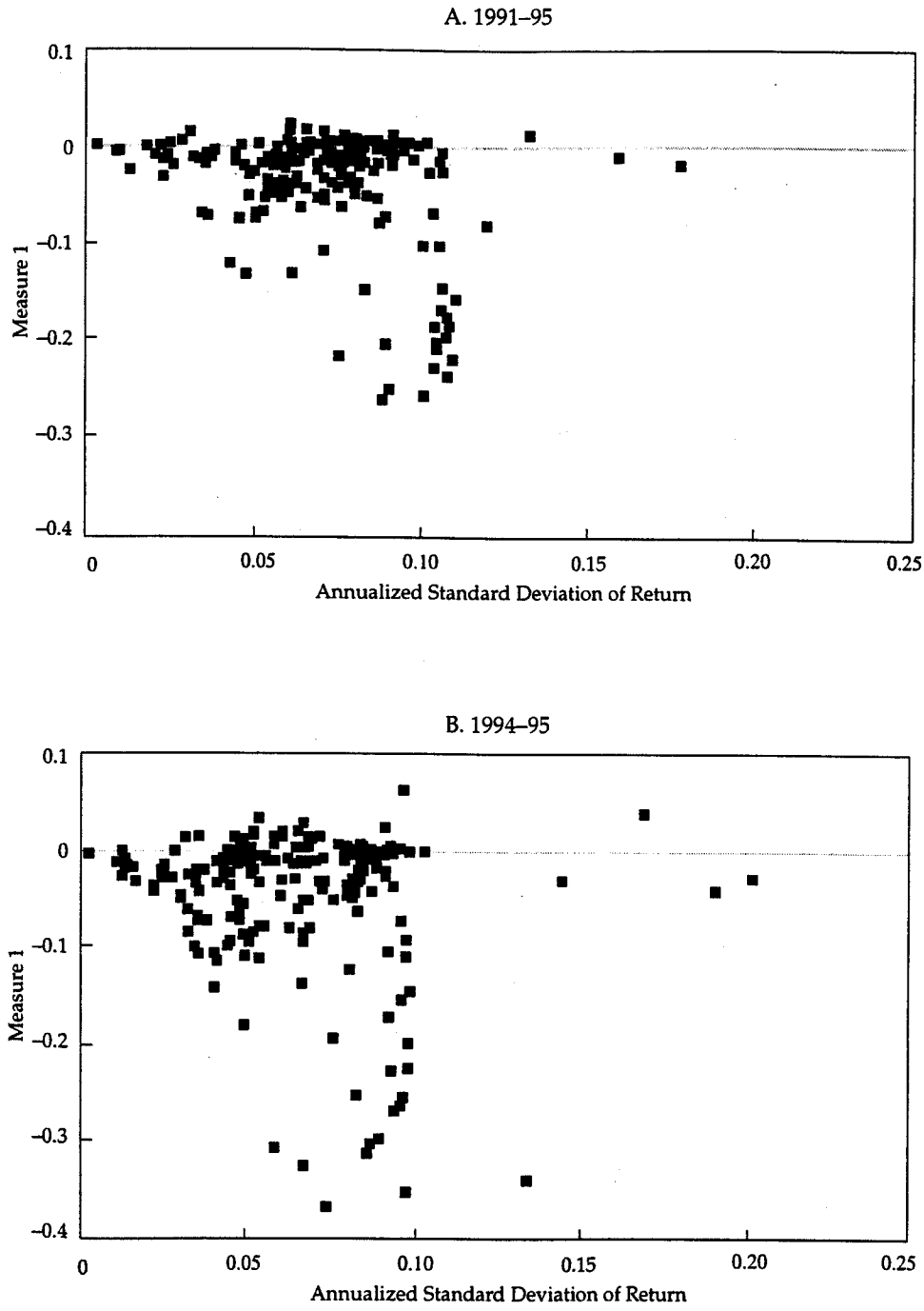



*Notes*: Each square represents the average annual return and standard deviation of return for an individual newsletter. The line plots out the average annual return and standard deviation of return for the efficient frontier—that is, for constant-weight portfolios consisting of a 100 percent T-bill contract and a 100 percent S&P 500 futures contract. The other points on the efficient frontier show the return and standard deviation for various combinations of T-bills and the S&P 500.

S&P 500, the unlevered fund return is greater than the buy-and-hold S&P 500 and the performance measure is positive. For Fund B, investors would have been better off acting on the newsletter recommendations than on a buy-and-hold strategy.

Both measures provide different perspectives, which can be seen in Figure 3 and Figure 4. For the evaluation period, Measure 1 draws an efficient

**Figure 2. Graham–Harvey Measure 1: Levering or Unlevering the S&P 500 to Achieve the Same Standard Deviation as the Fund**




*Note*: Each square represents the value of GH1 and the annualized standard deviation of return for an individual newsletter.

frontier using the S&P 500 and cash and checks to see if the newsletter lies above or below this constructed frontier. The volatility-matching approach inherent in GH1 is cleaner than the graphical analysis in Figure 1 because the newsletter return is compared with the return for a volatility-matched benchmark *over the exact same sample period.* Measure 2 compares all funds with a common level of volatility—the S&P 500 buy-and-hold volatility. All funds are on the same footing with GH2 and thus can be compared with each other.[6] The only potential disadvantage of GH2 is that it assumes the investor has the ability to lever an investment newsletter return to have the same volatility as the market.[7]

■ *New versus traditional performance measures.* How are our new performance measures related to

**Figure 3. Graham–Harvey Measure 2: Levering or Unlevering the Fund to Achieve the Same Standard Deviation as the S&P 500**

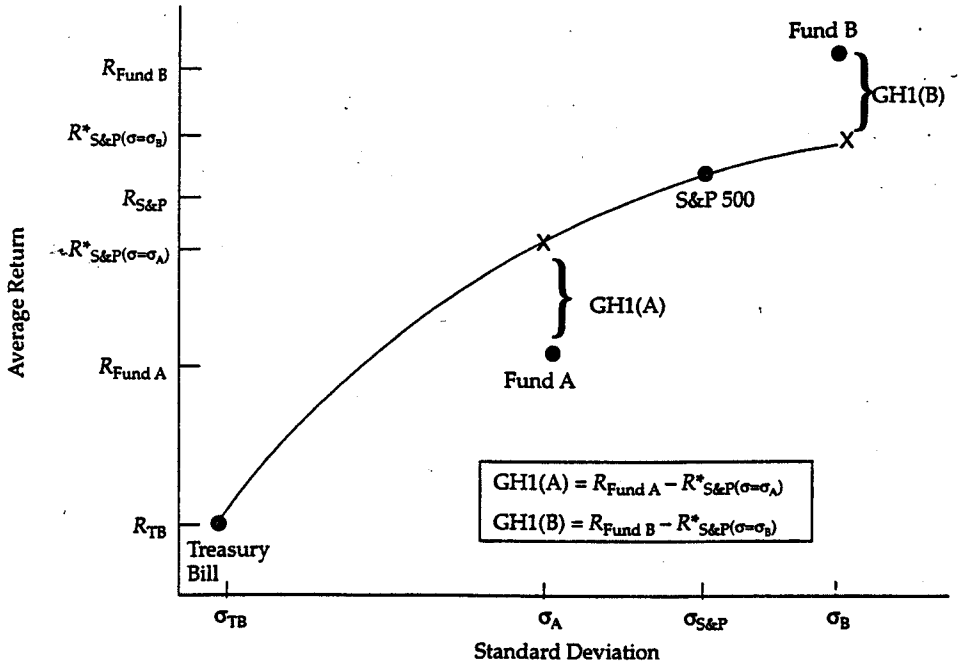
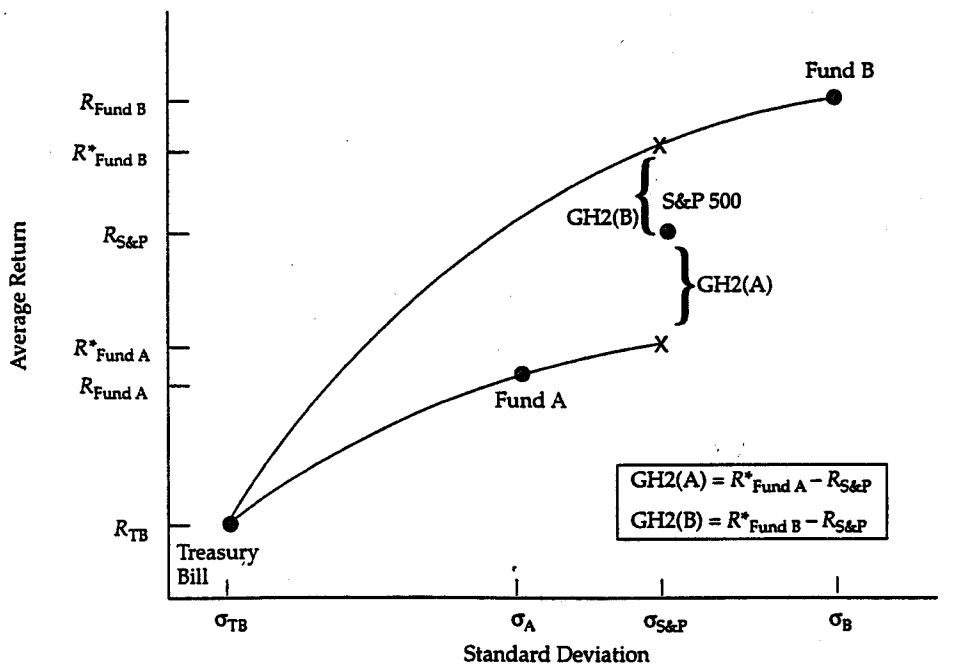



**Figure 4. Measure 1 and Annualized Standard Deviation of Return**




traditional measures? Consider the alpha from Sharpe's (1964) capital asset pricing model (CAPM). In the CAPM environment, the newsletter excess return is regressed on the market excess return. Roughly, the beta picks up the average level of market exposure. The alpha represents the extra return the newsletter earns over and above a position with a (fixed) average market exposure. This formulation is closely related to that of GH1, in which the market variance is adjusted to have the same variance as the newsletter. In GH1, however, the benchmark (market index and cash) will be constructed to have exactly the same volatility as the newsletter fund. In the CAPM, the benchmark

portfolio (beta times the market index) will have a different volatility from the fund. Using the CAPM, the newsletter fund volatility equals beta times the standard deviation of the market index return (the benchmark) *plus* the standard deviation of the idiosyncratic return. In contrast, GH1 exactly matches the total volatility of the newsletter fund.

To see the difference from another perspective, suppose a newsletter fund has a purely random strategy that switches between 200 percent long in the market and 200 percent short in the market. Also, suppose that the return from this random strategy happens to be 1 percent above the risk-free rate. If the CAPM beta is zero, then the alpha is 1 percent and this strategy would be identified as superior because it outearns the return implied by the CAPM. In contrast, GH1 would find a portfolio of S&P 500 futures and cash that has identical realized variance. This strategy would likely have twice the variance of the market. Hence, the random strategy would be compared with a buy-and-hold portfolio with double the variance of the market. Given that it only outearned the risk-free rate by 1 percent and yet doubled the market's volatility, the random strategy would be a significant underperformer according to the GH1 measure.

We calculated both GH1 and GH2 for all of the newsletters in our sample for the 1991–95 and the 1994–95 periods.[8] Interestingly, the overall performance of the newsletters is broadly consistent with Figure 1. Only 17.6 percent of the newsletters achieve a return greater than the volatility-matched fixed-weight portfolio (see Figure 4). This compares with the 37.0 percent of newsletters that lie above the curve in the informal analysis in Figure 1. In addition, only 24.1 percent of the volatility-adjusted newsletter portfolios achieve a return greater than the S&P 500 Index (using GH2 as a measure).

*An economic interpretation of the Graham–Harvey performance measures.* Our performance measures focus on long-run performance, and market timing is directly linked to long-run performance. In particular, GH1 compares the returns on the newsletter portfolios, whose weights change through time, with the returns on a constant-weight portfolio with equal volatility. It consists of two components, each of which has a direct link to market timing: (1) covariance between equity weights and market returns, and (2) a factor that penalizes changes in equity weights that do not time the market.

The idea of market timing is to reduce equity exposure before market declines and to increase exposure before market rallies. The successful timer's average return should be greater than the return on the constant-weight portfolio. Indeed, ignoring the cash returns, the following expression should be positive for a successful market timer:

$$E[w_i\, r_m] - E[w_i]E[r_m], \tag{1}$$

where $w_i$ represents newsletter $i$'s equity weights and $r_m$ is the market equity return. The first term is the average newsletter performance when $w_i$ is changing through time as recommendations change. The second term represents a return on a constant-weight strategy; the constant is the newsletter's average market exposure. Equation 1 is the definition of the covariance between weights and market returns. A positive covariance defines successful market timing—weights increase (decrease) during market rallies (declines). By definition, a positive covariance implies that the variable-weight newsletter strategy has a higher average return than the constant-weight strategy. Thus, the component of GH1 that measures the covariance between equity weights and market returns is a direct measure of market timing.

The second component of GH1 penalizes newsletters for changes in equity weights that do not time the market. To see why this result makes sense, notice that the variance of a newsletter's returns has two sources (when returns and weights are uncorrelated): the variance of equity returns and the variance of the weights. A newsletter that randomly changes weights induces volatility into its portfolio returns simply by changing the equity weights. The component of GH1 that penalizes strategies that are changing weights for the wrong reasons essentially says, "If you are changing weights and given that it is obvious that random weight changes contribute to variance, then you had better be changing weights to achieve a higher return; that is, you had better be timing the market."

What could cause a newsletter to change equity weights? Given its level of risk aversion, weights would change if (1) the newsletter believes expected returns are time-varying and/or (2) the newsletter believes market volatilities are time-varying. With time-varying expected returns, for example, the newsletter would increase (decrease) weights when the expected market return is above (below) the average expected return. A random shift in the weight may increase volatility, but a carefully planned shift in the weights to time the market may not. Further, expected returns should increase if the investor has some ability to detect time-varying expected market returns. That is, if the weights are changed in a way that successfully times the market, then average returns could increase and perhaps even variance could decrease.

In terms of mean–variance analysis, successful market timers should be above the efficient frontier

upturns and decrease market exposure during downturns. Remember that the mean–variance frontier is traditionally drawn with fixed investment weights. So even though changing weights are contributing to variance, the positive covariance between the weights and the market returns could actually decrease newsletter portfolio volatility for successful market timers. Thus, a successful market timer will have a positive value for GH1, which indicates a position above the efficient frontier. The successful timer will lie above the frontier because of positive covariance between the equity weights and the market and/or because changes in equity weights that do not time the market are not unduly penalized.

Our analysis of individual newsletter performance reveals a number of interesting insights. First, the range of average newsletter performance is striking. One of the highest profile newsletters, the *Granville Market Letter—Traders Portfolio* lost 2.2 percent a year during the past 13 years. The *Elliott Wave Theorist—Traders* lost an average of 10.1 percent a year since December 1985. Some performances have been impressive. *Medical Technology Stock Letter* produced annual returns of 24.8 percent a year from December 1985. The *Fidelity Monitor* produced 20.2 percent a year since December 1986. A large group of letters introduced in 1995 produced more than 30 percent in average annual returns (see Figure 1). As noted earlier, however, the average return on the S&P 500 futures plus cash in 1995 was 41 percent.

The equally weighted average newsletter return is 12.0 percent with 11.9 percent volatility from 1983 to 1995. A constant-weight portfolio of produced a 16.8 percent average annual return. Thus, the results for the equally weighted portfolio are consistent with the pairwise comparisons: The long-run performance of the equally weighted newsletter investment strategies is lower than a passive strategy with the same volatility.

## The Performance Report Card

Our performance "grade" is based on the sum of GH1 and GH2. Newsletters with performance in the top 10 percent received a grade of A. The second and third deciles were given a B, the fourth and fifth deciles were assigned a C, the sixth and seventh were designated D, the eighth and ninth got Es; the bottom 10 percent got the failing grade of F. Within each of the passing categories, we assigned plusses and minuses (highest third a plus and lowest third a minus). This grade distribution is tougher than most of those in school—our ranking scheme does not permit grade inflation.

Table 1 details the distribution of grades, the average GH1, and the average GH2 over the 1991–95 period. Note that both the median (Grade C–) measures are negative. The best newsletters (A+) produce annual returns 3.5 percent above the volatility-matched benchmark portfolio. The worst letters (F) produce annual returns 20.7 percent below the benchmark.

For comparison, we also included the Sharpe ratio of performance. This metric is the excess return on the newsletter divided by its volatility. In the mean–standard deviation graph (Figure 1), the Sharpe ratio is the slope of a line originating at the

**Table 1. Distributions of the Graham–Harvey Measures and Sharpe Ratios, 1991–95**

| | | Mean Values | | |
|---|---|---|---|---|
| Letter Grade | Number of Newsletters | GH1 | GH2 | Sharpe Ratio |
| A+ | 7 | 0.035 | 0.012 | 1.68 |
| A | 7 | 0.010 | 0.008 | 1.40 |
| A– | 7 | 0.006 | 0.004 | 1.45 |
| B+ | 14 | 0.004 | 0.002 | 1.41 |
| B | 15 | 0.001 | 0.000 | 1.43 |
| B– | 14 | –0.005 | –0.004 | 1.39 |
| C+ | 14 | –0.009 | –0.008 | 1.30 |
| C | 15 | –0.013 | –0.011 | 1.30 |
| C– | 14 | –0.020 | –0.013 | 1.04 |
| D+ | 14 | –0.026 | –0.019 | 1.14 |
| D | 15 | –0.039 | –0.021 | 1.01 |
| D– | 14 | –0.057 | –0.035 | 0.80 |
| E+ | 14 | –0.074 | –0.043 | 0.62 |
| E | 15 | –0.098 | –0.047 | 0.49 |
| E– | 14 | –0.137 | –0.072 | 0.10 |
| F | 22 | –0.207 | –0.203 | –0.72 |

risk-free rate and passing through the average annual return of a particular newsletter. A high Sharpe ratio means investors are getting more average return per unit of volatility than they would with lower ratios.[9]

Although the Sharpe ratio is a useful metric, it does not reveal the same type of information as GH1. In particular, the Sharpe ratio does not tell us what an investor could have achieved. In other words, a Sharpe ratio is hard to evaluate without a reference point. Roughly speaking, GH1 is related to the difference between the Sharpe ratio of the newsletter and the Sharpe ratio of the market. (The GH measures, however, differ from the Sharpe ratios in that they account for the curvature in the efficient frontier and make comparisons based on matched volatility.) Two newsletters can have the same Sharpe ratio, even if one lies above the efficient frontier and the other below it. In contrast, GH1 always assigns a positive score to a newsletter lying above the frontier and a negative score to one lying below it. Given that the efficient frontier is usually thought of as the dividing line between good and bad performers, the ability to make this distinction is a very desirable property. (GH2 has an analogous property.)

The average Sharpe ratios presented in Table 1 are correlated with the average GH1 and GH2. Indeed, the rank-order correlation for the 1991–95 period, reported in Table 2, exceeds 90 percent. GH1 and GH2, however, contain unique information. The rank-order correlation between the Sharpe ratio and Measure 1 (Measure 2) is 0.52 (0.59) for the 1983–90 period.

## THE PERFORMANCE GRADES AND FUTURE PERFORMANCE

Based on the GH1/GH2 letter grades during the 1986–90 period, we tracked the performance of the newsletters' portfolios for the next five years. Table 3 shows that a portfolio of all the A newsletters from 1986 to 1990 produced average returns of 12.6 percent in the next five years. The portfolio of newsletters that had an E or an F during the early period had only a 7.2 percent return in the next five-year period. The spread between highest and lowest was 5.4 percent on an annual basis.

Next, we calculated the letter grades based solely on the Sharpe ratios. The results are shown in Table 4. The Sharpe ratio shows less dispersion than the Measure 1/Measure 2 grades. The highest-rated letters delivered 12.7 percent annual return in the 1991–95 period and the lowest-rated letters produced 9.8 percent—a difference of 2.9 percent on an annual basis.

This result is consistent with the Hulbert (1995) study of 326 mutual funds. Using our Measure 1, Hulbert was able to identify a collection of funds that outearn the market *and* have lower volatility. Also, Hulbert's funds dominate those selected by using just the Sharpe ratio. This evidence suggests that performance is persistent, and our measures appear to be useful in assessing future performance based on past performance.

Hulbert was able to identify superior absolute performance. In our study, overall newsletter market-timing ability is so poor that we cannot identify superior absolute performance; however, our grading scheme does correctly rank groups of newsletters and identifies superior relative performance (e.g., the best relative to the worst). An investment in the A newsletters' strategies in December 1990 would have produced an annual return of 12.6 percent by December 1995 (see Table 3), but a passive strategy with the same volatility would have delivered 16.0 percent return. The important point is that our grading scheme identifies the best group of newsletters and hence provides valuable information. The evidence suggests that our measures can be used to identify superior absolute performers in a sample that includes good absolute performers. Overall, however, the average performance of the investment newsletter strategies is not particularly distinguished.

### Direct Measures of Market Timing

We also studied the movements in the S&P 500 futures index following changes in newsletter recommendations. The results, shown in Table 5, indicate that across all observations, the average S&P 500 excess return was 14.3 percent and was positive 71.8 percent of the time during the 1983–95 period.

**Table 2. Spearman Rank Correlations**

| Measure | GH Score | GH1 | GH2 | Sharpe Ratio |
|---|---|---|---|---|
| GH score | | 0.944* | 0.991* | 0.949* |
| GH1 | 0.986* | | 0.933* | 0.849* |
| GH2 | 0.993* | 0.962* | | 0.912* |
| Sharpe ratio | 0.563* | 0.520* | 0.593* | |

*Note*: The rank correlation coefficients for 1991 to 1995 appear above the diagonal; the rank correlation coefficients for 1983 to 1990 appear below the diagonal.

*Statistically significant at the 1 percent level.

**Table 3. Ability of the Graham–Harvey Grades to Predict Future Performance**

| GH Letter Grade (1986–90) | Mean Raw Return[a] (1991–95) | GH Score[b] (1991–95) | Sharpe Ratio (1991–95) |
|---|---|---|---|
| A | 12.6% | –0.034 | 1.26 |
| B | 11.6 | –0.046 | 1.04 |
| C | 11.6 | –0.058 | 1.00 |
| D | 11.4 | –0.083 | 0.92 |
| E or F | 7.2 | –0.215 | 0.40 |
| A,B,C | 11.8 | –0.049 | 0.62 |
| D,E,F | 8.9 | –0.157 | 1.07 |

*Note:* A newsletter must make a recommendation in at least 30 months during each five-year period (1986–90, 1991–95) to be included in this analysis.
[a]Annual return earned by following a newsletter's market-timing recommendations.
[b]Sum of GH1 and GH2.

**Table 4. Ability of the Sharpe Grades to Predict Future Performance**

| Sharpe Letter Grade[a] (1986–90) | Mean Raw Return (1991–95) | GH Score[b] (1991–95) | Sharpe Ratio (1991–95) |
|---|---|---|---|
| A | 12.7% | –0.038 | 1.01 |
| B | 12.4 | –0.042 | 1.17 |
| C | 11.8 | –0.069 | 1.02 |
| D | 10.9 | –0.103 | 0.85 |
| E or F | 9.8 | –0.180 | 0.74 |
| A,B,C | 12.2 | –0.054 | 0.78 |
| D,E,F | 10.2 | –0.146 | 1.08 |

*Note:* A newsletter must make a recommendation in at least 30 months during each five-year period (1991–95) to be included in this analysis.
[a]Newsletters that have a Sharpe ratio in the top decile are assigned a Sharpe grade of A, those in the second and third deciles are assigned a B, those in the fourth and fifth deciles are assigned a C, those in the sixth and seventh deciles are assigned a D, those in the eighth and ninth deciles are assigned an E, and those in the lowest decile are assigned an F.
[b]Sum of GH1 and GH2.

First, we examined the market return in the month after a recommended increase in equity weight. If the newsletters have market-timing ability, the return after an increase should be greater than the overall average. But it is not. The average return after increased equity weights is 13.7 percent and is positive 71.3 percent of the time. The market performance after decreased equity weights is better. This result is exactly the opposite of what one would expect from successful market timing.

Perhaps the investment newsletters have a longer than one-month horizon return in mind when they change weights. Table 5 shows the six-month S&P 500 excess return following increased and decreased weights. The story is the same as for the one-month returns. After equity weight increases, the average annualized six-month return is 12.2 percent, and it is 15.7 percent when weights decrease. In addition, the six-month excess return is positive 70.3 percent of the time after equity weight increases compared with 72.9 percent after equity weight decreases.

We also examined the performance of the newsletter recommendations preceding big movements in the market—that is, those with absolute values exceeding one standard deviation. The performance was only slightly better than in the one- and six-month analysis. After equity weight increases, the average large movement in the market return is 30.3 percent. After equity weight decreases, the average large movement in the market is 27.4 percent. Among the large movements, the market rose 67.6 percent of the time after weight increases and 68.3 percent of the time after weight decreases.

## Hot Hands and Cold Hands

Is there a "hot hands" phenomenon in the newsletter recommendations? That is, if the newsletter produces a correct recommendation, does this imply that the next recommendation has a higher chance of being correct? According to Table 5, if the previous recommendation was correct (a one-recommendation hot streak), the average market return after the next market weight increase is 16.4 percent and the market return after the next

| | Next Month's S&P 500 Excess Return | | Next Six Month's S&P 500 Excess Return | | Large Movements in Next Month's S&P 500 Excess Return | |
|---|---|---|---|---|---|---|
| | % > 0 | Mean | % > 0 | Mean | % > 0 | Mean |
| *All observations:* | 71.8% | 0.143 | 71.6% | 0.140 | 68.0% | 0.287 |
| Observations in which equity weights | | | | | | |
| Increased ($\Delta w_t > 0$) | 71.3 | 0.137 | 70.3 | 0.122 | 67.6 | 0.303 |
| Decreased ($\Delta w_t < 0$) | 72.3 | 0.149 | 72.9 | 0.157 | 68.3 | 0.274 |
| ($p$-Value)[a] | (0.344) | (0.364) | (0.496) | (0.499) | (0.382) | (0.293) |
| *One-recommendation hot streak*[b] Observations in which equity weights | | | | | | |
| Increased ($\Delta w_t > 0$) | 74.6 | 0.164 | 72.8 | 0.145 | 70.0 | 0.342 |
| Decreased ($\Delta w_t < 0$) | 68.8 | 0.122 | 70.6 | 0.132 | 66.3 | 0.219 |
| ($p$-Value)[a] | (0.001) | (0.008) | (0.067) | (0.195) | (0.136) | (0.075) |
| *One-recommendation cold streak*[c] Observations in which equity weights | | | | | | |
| Increased ($\Delta w_t > 0$) | 69.0 | 0.117 | 68.4 | 0.105 | 66.1 | 0.278 |
| Decreased ($\Delta w_t < 0$) | 73.3 | 0.141 | 72.8 | 0.152 | 66.0 | 0.232 |
| ($p$-Value)[d] | (0.011) | (0.067) | (0.013) | (0.002) | (0.487) | (0.263) |
| *Three-recommendation hot streak*[e] Observations in which equity weights | | | | | | |
| Increased ($\Delta w_t > 0$) | 76.4 | 0.177 | 73.6 | 0.172 | 72.3 | 0.342 |
| Decreased ($\Delta w_t < 0$) | 65.7 | 0.043 | 68.2 | 0.074 | 53.5 | –0.104 |
| ($p$-Value)[a] | (0.001) | (0.001) | (0.035) | (0.017) | (0.006) | (0.004) |
| *Three-recommendation cold streak*[f] Observations in which equity weights | | | | | | |
| Increased ($\Delta w_t > 0$) | 67.2 | 0.088 | 67.2 | 0.095 | 61.5 | 0.240 |
| Decreased ($\Delta w_t < 0$) | 72.7 | 0.126 | 71.0 | 0.125 | 67.7 | 0.351 |
| ($p$-Value)[d] | (0.031) | (0.096) | (0.103) | (0.117) | (0.224) | (0.263) |

*Notes*: The top portion of the table reports on market movements after increases ($\Delta w_t > 0$; see row 2) and decreases ($\Delta w_t < 0$; see row 3) in recommended equity weights. The middle section explores market movements conditional on whether a newsletter correctly called the direction of the market in its last recommendation. The bottom section examines market movements after recommendations by newsletters that have a "hot hand" (i.e., they correctly anticipated the direction of the market in their last three recommendations) versus letters that have a "cold hand" (i.e., they incorrectly anticipated the direction of the market in their last three recommendations).

[a]$p$-Value for a one-tailed ANOVA $F$-test testing the null hypothesis that the mean values in the two rows above are equal against the alternative hypothesis that the value associated with $\Delta w_t > 0$ is greater than the value for $\Delta w_t < 0$. A value of 0.05 or smaller indicates that the null is rejected in favor of the alternative at a 5 percent level of significance.

[b]A one-month hot streak occurs when the newsletter's previous recommendation correctly anticipated the direction of the movement in the S&P 500 futures contract.

[c]A one-month cold streak occurs when the newsletter's previous recommendation did not correctly anticipate the direction of the movement in the S&P 500 futures contract.

[d]$p$-Value for a one-tailed ANOVA $F$-test testing the null hypothesis that the mean values in the two rows above are equal against the alternative hypothesis that the value associated with $\Delta w_t < 0$ is greater than the value for $\Delta w_t > 0$.

[e]A three-month hot streak occurs when the newsletter's previous three recommendations all correctly anticipated the direction of the movement in the S&P 500 futures contract.

[f]A three-month cold streak occurs when the newsletter's previous three recommendations did not correctly anticipate the direction of the movement in the S&P 500 futures contract.

weight decrease is only 12.2 percent. After equity weight increases, 74.6 percent of the next month's returns are positive compared with 68.8 percent when equity weights recommended decreases. Similar but weaker results are evident when the equity return is measured over six months. The hot-hands phenomenon is also present in the large return movements. After equity weight increases, the average large return is 34.2 percent, compared with 21.9 percent when weights decreases are rec-

ommended.

Symmetric results are found for cold hands. If the previous recommendation was incorrect, the average equity return after the next recommended weight increase was 11.7 percent compared with 14.1 percent following a recommended weight decrease. Even stronger results were found for the six-month returns. Apparently, investors should take the opposite investment strategy from that recommended by the newsletters with cold hands.

Table 5 also reports on a hot (cold) streak of three consecutive correct (incorrect) recommendations. After recommended increases, the average market excess return was 17.7 percent and positive 76.4 percent of the time. After recommended decreases, the average excess return was only 4.3 percent and positive 65.7 percent of the time. Similar results were found for returns over the next six months. In addition, the hot hands were able to position for large market movements. The average large market return after weight increases was 34.2 percent compared with –10.4 percent when decreased weights were recommended.

The cold-hands phenomenon was also strongest for three incorrect recommendations. The average return after equity weight increases was 8.8 percent (positive 67.2 percent of the time), and it was 12.6 percent (positive 72.7 percent of the time) after recommended equity weight decreases. The same pattern emerges for the six-month returns, as well as the large movements in the market. As was the case for the single-recommendation cold streak, investors would be better off implementing a strategy opposite from that recommended by the cold-hand newsletters.

Overall, a significant hot-hands/cold-hands effect is present in the newsletter recommendations. To implement this strategy, however, one needs to collect all of the investment newsletters. Furthermore, the evidence based on a smaller data set in Graham and Harvey (1994, 1996) suggests that the hot-hands phenomenon cannot be used to select a newsletter with superior long-run performance. In Monte Carlo simulations, Graham and Harvey found that only 11 of 237 newsletter strategies they examined could be deemed superior in the long run. This number is fewer than one would expect by pure chance;[10] at Graham and Harvey's 10 percent significance level, the simulations would be expected to identify 24 strategies that would appear "superior" even if the newsletters made completely random recommendations. Hence, the hot-hands phenomenon has economic implications across the range of newsletters, not for a particular newsletter, and appears to reverse itself in the long run for any given newsletter (e.g., a hot-hand letter eventually develops a cold hand).

## CONCLUSIONS

We have proposed two measures for evaluating the performance of asset-allocation recommendations from investment newsletters. The first measure compares the newsletter's portfolio return with a portfolio of S&P 500 futures and cash that has the same volatility over the evaluation period. The benchmark has fixed investment weights, and the newsletter strategy has variable weights. Presumably, if the newsletter is successfully timing the market (increases weights before market upticks and decreases weights before market downticks), the newsletter should be able to outperform this passive benchmark.

A second measure adjusts the newsletter's volatility strategy. We constructed a portfolio of the newsletter strategy and a Treasury bill with exactly the same volatility as the S&P 500. The difference between the returns on the volatility-adjusted strategy and the S&P 500 defines Measure 2.

Using a performance report card for 326 newsletter strategies, we evaluated those strategies over the past five years and the past two years. Overall, the performance is unimpressive. Indeed, some high-profile newsletters have performed remarkably poorly over long periods of time. Our performance measures, however, were able to identify the best newsletters. We also presented evidence that our new measures are superior to the Sharpe ratio.

We found that newsletter recommendations contain important information based on their past performance. Recommendations of the hot-hand newsletters have some ability to predict up and down movements in the market. The cold-hand newsletters also provide important information. Investors would be better off ignoring or investing opposite to their advice.

Importantly, to implement a trading strategy based on the hot-hands/cold-hands phenomenon, one has to subscribe to all the investment newsletters. Given that the typical newsletter in our sample cost $200 for an annual subscription, this endeavor could be expensive for an individual investor. The evidence suggests that the hot-hands phenomenon cannot be used to identify a particular superior newsletter over the long run—the hot-hands phenomenon is fleeting.[11]

## NOTES

1. More information on our performance metrics can be found at http://www.duke.edu/~charvey/performance_eval/lettab1.htm.
2. We conducted a sensitivity analysis by excluding those newsletters with broader allocations and found that they do not alter our results.
3. Hulbert may "clarify" the data if a newsletter is vague in its recommendations. For example, if a newsletter is 50 percent long the market and 50 percent in cash but recommends a one-month hedge against the long, Hulbert may impute a 50 percent short in futures with a 50 percent margin; this method hedges the long position but avoids transaction costs for closing out the long position. For consistency among our data, however, in situations where both long and short positions are greater than zero, we took the net position and assigned the remainder to cash.
4. This strategy is passive in the sense that investment weights do not change through time. Monthly rebalancing is necessary, however, to maintain the fixed weights.
5. The means and standard deviations of the hypothetical portfolios' returns for various horizons are available on request. Also, newsletter-specific results are available for 292 strategies with the clearest domestic equity/cash recommendations.
6. Modigliani and Modigliani (1997) applied a measure similar to GH2 to a sample of mutual funds, but they did not allow for curvature in the efficient frontier (see their Exhibit 1). That is, they assumed that the cash return has zero variance and zero covariance with other assets. This assumption is true only if the maturity of the cash instrument exactly coincides with the evaluation period. Further relative to GH2, the assumption could result in misleading inferences about the performance of low-volaility funds that need substantial leverage to achieve the S&P 500 volatility.
7. In our sample of market timers, the extreme use of leverage is not an issue, but in applying GH2 to a broader class of asset returns, such as mutual funds, this issue could be a problem. For example, substantial leverage would have to be used to lever a money-market fund to achieve the volatility of the S&P 500.
8. These results are available from the authors on request.
9. The Sharpe ratios of the newsletters over the past five years and the past two years are also available on request.
10. Of the 11 letters that Graham and Harvey (1996) found "superior" in the period ending in December 1992, only three had a positive GH1 in the January 1993 to December 1995 period: *Investor's Guide/Closed-End Funds IV, Nurock TMI* (no short), and *Investor's Intelligence-Long Term.* One letter had a zero GH1: *Switch Fund-Conservative/Momentum.* Over the January 1994 to December 1995 period, only a single letter, *Investor's Guide/ Closed-End Funds IV,* had a positive GH1. This performance confirms Graham and Harvey's intuition that care must be taken in deeming a strategy "superior."
11. We thank Mark Hulbert of the *Hulbert Financial Digest* for providing us with the data and answering many questions.

## REFERENCES


Graham, John R., and Campbell R. Harvey. 1994. "Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendations." National Bureau of Economic Research Working Paper 5597 (October).

———. 1996. "Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendations." *Journal of Financial Economics,* vol. 42, no. 3 (November):397–422.

Hulbert, Mark. 1995. "The Graham–Harvey Test." *Forbes* (June 19):160–61.

Modigliani, Franco, and Leah Modigliani. 1997. "Risk-Adjusted Performance." *Journal of Portfolio Management,* vol. 23, no. 2 (Winter):45–54.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance,* vol. 19, no. 3 (July):425–42.