Evaluating Bond Fund Sector Timing Skill

George Comer*
Georgetown University
August 2005

*Contact information: 417 Old North, Georgetown University, Washington, DC, 20057. Phone: 202-687-0676, Email: gc45@georgetown.edu. I acknowledge support from the Robert H. Steers Faculty Research Fellowship provided by the McDonough School of Business.
Abstract

Reported portfolio data indicate that bond fund managers engage in sector timing behavior. I use simulation procedures to examine the ability of various versions of the Treynor and Mazuy (1966) timing specification to detect positive sector timing skill. Results indicate that the models are unable to detect timing ability at reasonable skill levels for the majority of managers. Alternative measures which compare actual fund returns to the returns from passive strategies using the previous period’s sector weights are better able to detect positive ability at reasonable skill levels. When applied to a sample of general government bond funds, these measures indicate managers may possess positive sector timing ability.
I. Introduction

The academic literature has devoted substantial effort to measuring the market timing ability of portfolio managers. Often, researchers rely on the returns based models first developed by Treynor and Mazuy (1966) and Henriksson and Merton (1981) to estimate market timing skill. In their original forms, these models examine a manager’s ability to switch assets between a broad market portfolio of stocks and Treasury bills. Timing skill is measured by the sign and statistical significance of a regression coefficient designed to estimate portfolio shifts made by the manager.

An interesting empirical issue is whether these returns based models and their multi-index extensions are able to accurately measure timing ability when a fund explicitly engages in sector timing behavior. This question is motivated by two observations. Available annual portfolio data indicate that most bond fund managers not only engage in traditional market timing but also shift assets across the major sectors (government, corporate, mortgage, high yield, and foreign) of the fixed income market in an attempt to generate higher returns. The return distributions of these sectors often differ providing an opportunity for a skilled manager to generate higher returns than a manager with no timing ability. Given the growing importance of bond funds in many investors’ portfolios, the ability to correctly identify superior managers is of great interest.\textsuperscript{1}

Secondly, Kothari and Warner (2001) examine mutual fund performance measures using standard risk adjustment techniques which are well established in the literature. Using both the Fama-French three factor model and Carhart four factor model, they find that these measures have little ability to detect large magnitudes of abnormal fund performance, especially if a fund’s characteristics differ from those of the market portfolio. They suggest that procedures which
make use of portfolio holdings may have more power than traditional returns based methods. Similarly, when examining the timing skill of equity fund managers, portfolio based timing measures (see Jiang, Yao, and Yu (2004)) appear to be a preferable method to the returns based measures of timing.

However, comparable portfolio data is not available for bond funds rendering such techniques difficult to implement. Thus, studies examining the timing skill of bond funds (and hybrid or global funds which hold a significant portion of bonds in addition to stocks) are likely to rely on variations of traditional returns based models. Two of the first studies to examine the timing ability of bond funds both infer skill from the fund’s return series. Chen, Ferson, and Peters (2005) use a variation of the Treynor Mazuy quadratic approach while Comer, Boney, and Kelly (2005) combine the use of Sharpe (1992) style analysis with the Henriksson and Merton methodology.²

Thus, the goal of this study is to use simulation procedures to focus on the ability of the Treynor Mazuy quadratic regression approach to identify timing skill when managers engage in timing across multiple bond sectors. Simulated returns are generated for portfolio strategies which are designed to mimic reported annual portfolio data provided by government bond funds to Morningstar. Model power is examined along several dimensions. One dimension is manager skill level. A second dimension is the benchmark used to measure skill. A third dimension is the version of the Treynor Mazuy (hereafter referred to as TM) model used to measure performance. Several variations of the quadratic approach have been suggested, and I examine three versions that have appeared in the literature.³

I examine the ability of the model to detect positive timing ability at reasonable skill levels. Results indicate that the various versions of the TM model are weak, and managers must
have substantial forecasting skill (i.e. forecasting accuracy of 80% and higher) before the models are likely to signal significant positive timing ability for a large number of managers. Given the typical portfolio strategies employed by managers and the return distributions of the bond sectors, it is difficult to generate sufficient convexity in returns such that the tests will detect significant timing skill for a large percentage of funds at reasonable skill levels.

I examine an alternative measure that I refer to as an attribution return. This measure compares actual monthly fund returns to estimated returns from passive strategies based on lagged sector weights. Attribution returns are calculated using both annual reported weights and estimated average monthly weights. This measure is better able to detect positive ability at reasonable skill levels (60%-80%). When applied to a sample of general government bond funds, the attribution return measure indicates that managers may possess positive sector timing ability.

This paper is organized as follows: Section II provides the details concerning the design of the simulation. Section III discusses the simulation results focusing on the size and power of the TM model. Section IV examines the performance of the attribution return measure and compares the results to the TM model. Section V applies the attribution return measure to the sample of government bond funds. Section VI concludes.

II. Structure of Simulation

To examine the TM model, I generate a time series of returns for 500 simulated managers with the goal of mimicking the characteristics of the returns of actual bond fund managers. In this section, I focus on the structure of the simulation. I define 1) the portfolio strategies used by managers, 2) the sample period, 3) managerial skill level, 4) the simulated return series, and 5)
versions of the TM model to be examined. The section ends with a brief discussion of the known problems of the quadratic regression approach.

A. Portfolio Strategies Used by Bond Fund Managers

Beginning with the March 31, 2005 CD, Morningstar Principia provides a history of the annual investments weights by sector for various categories of bond funds (general government, Treasury, general corporate, and high quality corporate) going back to 1998. Morningstar obtains this portfolio data from voluntary surveys sent out to the mutual funds within its database. The portfolio is divided into the following five sectors for which sector weights are available 1) US governments, 2) mortgage, 3) credit, 4) foreign, and 5) cash. The annual figures reflect the last data received from a fund for each calendar year which is usually December 31.

I focus on the data for general government bond funds. This is the second largest Morningstar category among bond funds with combined net assets of over $83 billion. Morningstar defines this category as funds that pursue income by investing in a combination of mortgage-backed securities, Treasuries, corporates, and agency securities. As made clear from this definition, these funds have the flexibility to engage in both market and sector timing in order to generate high returns. I collect the annual portfolio data for these funds to calculate the variation in portfolio weights which would suggest whether the managers engage in market and sector timing activity. If there is evidence of timing activity, the portfolio data will then indicate a typical portfolio strategy employed by this category of funds that could form the basis of the simulation.

Morningstar lists 129 general government bond funds that have been in existence over the entire 1998-2004 period. 122 of these funds are missing at most one set of annual weights.
Table 1 provides descriptive statistics of the reported portfolio weights. I combine the reported data for the government and credit sectors. As measured over this period by the returns from the Lehman government index and the Lehman credit index, these two sectors are highly correlated with a correlation value of 0.86. The test statistic for a matched pairs t-test is 0.69 indicating there is not a significant difference in the return distribution of the two sectors. For well diversified portfolios, positive timing skill across these two sectors would not generate a significant difference in returns when compared to perverse timing skill.

The average portfolio allocation is 53.7% government/credit, 37.5% mortgage, 7.8% cash, and 1.0% foreign. The overall portfolio allocation and the descriptive statistics are not skewed by a few outliers as the median portfolio allocation is virtually identical (54.6%/37.3%/6.8%/1.3%) to the mean. For each fund, I calculate the standard deviation of the portfolio weights for each of the sectors, and then calculate the average standard deviation across all funds. I calculate the range of portfolio weights in a similar manner.

The portfolio data provide strong evidence of sector timing activity over the 1998-2004 period. Because the average weight for the foreign sector is so small, I exclude that sector from the following calculations. The average standard deviation for each of the sectors ranges from 7.3% (cash) to 12.3% (government/credit) which indicates significant variation in the annual weights. The variation in weights is sufficiently large that it can not be attributed to the weights evolving due to a buy and hold strategy. The average range of sectors weights varies from 19.1% (cash) to 33.4% (government/credit). Evidence of timing activity is consistent across the entire fund sample. For over 85% of the funds, the range of reported mortgage and government/credit sector weights is greater than 10%, while over 70% of the funds demonstrate such variation in the cash sector weights.
Given this data, I define a simulated allocation strategy which reflects the behavior of these funds. Each simulated fund manager is assumed to have a neutral allocation for each sector of 55% government/credit, 35% mortgage, and 10% cash. A manager makes forecasts of the performance of each of the three sectors. If he expects a sector to be the best (worst) performing of the three, he will increase (decrease) his allocation to that sector by 10% from the neutral allocation. If the sector is not expected to be the best or worst performing, the manager maintains the neutral allocation for that sector. For example, let $R_j$ represent the total return for sector $j$. For month $t$, if a manager forecasts $R_{gc,t} > R_{mort,t} > R_{cash,t}$, then his allocation would be 65% government/credit, 35% mortgage, and 0% cash.

**B. Sample Period**

Simulated returns will be based on the distribution of the monthly bond sector returns over the time period 1998-2004. This seven year time period is primarily motivated by the access to the portfolio weight data described in the previous section. After examining the power of the TM model, I can compare the results to conclusions drawn from an alternative estimate of timing skill strictly based on the annual weights.

There are two issues concerning the use of this time period in a simulation. First, there needs to be sufficient variation in bond returns over the time period such that there are ample timing opportunities across the three sectors. Second, there needs to be sufficient evidence that the length of the time period is representative of the typical tenure of a fund manager.

To examine the first issue, I use the Lehman Brothers total return bond indices. The Lehman Brothers indices are the most comprehensive market weighted bond indices available and the most widely used family of indices for bond portfolio evaluation. For the purposes of
this study, the government/credit sector is represented by the Lehman government/credit index. Mortgages are represented by the Lehman mortgage backed securities index. Cash investments are represented by the Lehman Treasury bill index.

Over the seven year period, the government/credit sector has the highest average annual return of 6.8%. The mortgage sector follows with a 6.4% return, and cash has the lowest annual average of 3.7%. On an annual basis, each sector has the highest return for at least one year during the sample period. The government/credit sector has the highest return for 1998 and from 2000 through 2003. Cash performed best during 1999 with returns 2.8% greater than the mortgage sector and 6.9% greater than the government/credit sector. The mortgage sector has the highest return in 2004 with returns 0.5% greater than the government/credit sector and 3.0% greater than the cash sector.

On a monthly basis, the government/credit sector is the best performing sector during 37 of the 84 months which represents 44% of the time period. The cash and mortgage sectors have the best performance during 24 and 23 months (representing 29% and 27% of the months) respectively. Across the 24 months when cash has the highest return, the sector outperforms the mortgage sector by an average of 0.6% and the government/credit sector by 1.2%. During the 23 months when the mortgage sector is the best performing, the sector outperforms cash by an average of 0.5% and the government/credit sector by 0.3%. These are substantial differences when compared to the average monthly returns of 0.6% for the government/credit sector, 0.5% for the mortgage sector, and 0.3% for the cash sector.

Given that the evidence indicates there is ample opportunity for a skilled manager to generate superior returns through sector timing over this time period, I turn to the second issue of the typical tenure of a fund manager. Given the funds on which the simulation is based have
been in existence for at least seven years, tenure could be much greater than seven years if there have been no management changes made by any of the funds.

Morningstar reports the tenure of the manager for each of the 122 funds included in the sample. Manager tenure is defined as the number of years that the current manager has been the portfolio manager of the fund. For funds with more than one manager, the average tenure is reported. If there is only one manager, who has been at the fund for less than six months, Morningstar does not report manager tenure. If no data is reported, I assume the manager has been managing the fund for six months, thus my average is likely to be higher than the true estimate.

I find that the average manager tenure is 5.3 years and the median is 4.9 years. This is despite the fact that as of the end of 2004, the average fund in the sample has been in existence for 14 years. Only 35% of the current fund managers (43 of 122) have been on the job longer than seven years. Thus, the evidence indicates that the seven year period is representative of the typical tenure of a fund manager and appropriate for use in the simulation.6

C. Managerial Skill Level

No manager will be able to perfectly forecast the returns of the various sectors. Thus, I define the simulated manager’s skill level. Skill is defined as his ability to correctly forecast the ordering of the performance of the sectors which he is timing. Skill level varies from 50% (neutral timing ability) to 100% (perfect timing ability). I examine skill levels of 50% and above because I am focusing on the model’s ability to detect significant positive timing ability.

The simulation period of 1998-2004 covers 84 months. Each simulated dataset (to be defined in the next section) includes 84 observations. Each observation in a specific dataset is
assigned a number $x$ from 1 to 84 without replacement. Suppose I am simulating the returns of a manager with a 50% skill level. Then if a month is assigned a number less than 43, I assume that the simulated manager correctly forecasts the order of the performance of the three sectors. For simplicity, I define an incorrect forecast as the manager predicting the reverse order of the relative performance of the three sectors (i.e. he predicts the best sector as being the worse and vice versa).

D. Simulated Manager Returns

In order to implement the portfolio strategies across the various skill levels defined in the previous sections, I create 500 simulated data sets where each dataset is composed of a time series of 84 observations of the returns for each of the three sectors (government/credit, mortgage, and cash). The goal of the simulated data sets is to mimic the return characteristics of actual portfolios that are held by a general government bond fund manager.

One approach in creating the data sets involves generating monthly data for each of the sectors based on the distribution of the monthly returns of the sector. This approach is used by Goetzmann, Ingersoll, and Ivkovic (2000) where they generate daily excess returns of broadly diversified U.S. stock market indices. However, this approach is problematic for this simulation. The distributions of the three sectors indicate that none of them is normally distributed and that a multivariate normal distribution would not be appropriate. No other distribution is available that could reasonably fit the data and maintain both the time series and cross section properties of the returns for the three sectors.

A second approach is to construct a simulated return series for each sector by randomly selecting a specific number of individual bonds from that sector and calculating the time series
return of the portfolio of bonds. This approach is used by Kothari and Warner (2001) who use simulation procedures to examine the ability of various models to detect abnormal performance among equity funds. It is difficult to construct such a time series as the databases for individual bonds and their returns are not as developed as their stock counterparts. However, I have access to the full complement of Lehman bond indices and can use those to create the simulated return series for each sector.

The index for each of the three sectors is a weighted average of a set of subindices. For example, the government/credit index is composed of the Lehman 1-5 year government/credit index, the Lehman 5-10 year government/credit index, and the Lehman long (10+) government/credit index. Likewise, the mortgage index is composed of the GNMA index, the FHLMC index, and the FNMA index. The Treasury index is composed of the following four indices: 1) 1-3 month Treasury, 2) 3-6 month Treasury, 3) 6-9 month Treasury, and 4) 9-12 month Treasury. By using these subindices, I can create return series that reflect the returns generated by actual bond fund managers.

To create the simulated data sets for each sector $j$, I first generate 500 sets of random numbers, $w_i$, for each sector that satisfies the following constraints.

For the government/credit sector

$$0 < w_{gc^{0105}} \cdot w_{gc^{0510}} \cdot w_{gc^{long}} < 1 \quad and \quad w_{gc^{0105}} + w_{gc^{0510}} + w_{gc^{long}} = 1 \quad (1)$$

For the mortgage sector

$$0 < w_{gnma} \cdot w_{fhlmc} \cdot w_{fnma} < 1 \quad and \quad w_{gnma} + w_{fhlmc} + w_{fnma} = 1 \quad (2)$$

For the cash sector
Then, using the sets of 500 random weights and the actual monthly returns over the 84 month sample period for each subindex, I am able to create 500 simulated returns series for each of the three sectors. For example, the simulated return for the government/credit sector for manager $i$ during month $t$ is

$$R_{sgr.i} = w_{gc0103} \cdot R_{gc0103.t} + w_{gc0006} \cdot R_{gc0006.t} + w_{gc0009} \cdot R_{gc0009.t} + w_{gc0012} \cdot R_{gc0012.t}$$

Then the simulated fund return for manager $i$ for month $t$ is given by

$$R_{it} = w_{gc,i} \cdot R_{sgr,i} + w_{mort,i} \cdot R_{smort,i} + w_{cash,i} \cdot R_{scash,i}$$

where $w_{gc,i}$, $w_{mort,i}$, and $w_{cash,i}$ are based on the portfolio strategy and skill levels defined in the previous section. The advantage to this procedure is that each fund’s return series reflect sector asset weightings that deviate from the broad sector index. In a later section, I use the subsector weights to examine the impact that a fund’s style characteristics have on estimating timing ability.

I want to compare the return distribution of these simulated series to the returns of the fund sample. To facilitate comparison, I compute estimated monthly gross returns for each fund in the sample. Monthly fund returns are obtained from the CRSP Mutual Fund Database. However, the reported returns are net of expenses while the simulated series do not have expenses deducted. Each fund’s annual expense ratio is also reported by the CRSP Database. I divide the annual expense ratio by twelve to obtain an estimate of monthly expenses. I then add this amount to the reported fund return to obtain the estimated monthly gross return.

Table 2 reports descriptive statistics for the simulated and actual sample returns. Because I do not know the skill level of the real life managers, I report simulated return statistics for three skill levels: perverse timing ability (0% skill), neutral timing ability (50% skill), and perfect
timing ability (100% skill). I compute the statistics for each individual fund and then average the statistics across the funds to obtain the values reported in the table.

The results indicate that as skill level increases from perverse to perfect, all descriptive statistics change monotonically. More importantly, all statistics for the fund sample fall within the range of values represented by the 0% and 100% skill levels with the average standard deviation being the only exception. The mean and standard deviation of the fund sample (0.510% and 0.98% respectively) closely match the equivalent statistics for the neutral skill level (0.531% and 1.05%). The skewness and kurtosis of the sample (-0.571 and 1.260 respectively) most closely reflect the statistics of the perfect skill level (-0.500 and 1.029). The simulated data sets clearly mimic the return characteristics of actual portfolios that are held by a general government bond fund manager. I use these simulated returns in Section III to examine the power of the models that I introduce in the following section.

E. Models

I examine the power of three versions of the TM model that have been commonly used in the academic literature to measure timing ability. The first model is the original TM model that was developed in 1966. They examine timing ability by testing for nonlinearity in returns. They measure the following relationship between the excess return of the fund and the excess return of the market by running the following regression:

\[ r_{it} = a_i + b_1 r_{mkt,t} + \lambda_{i,mkt} r_{mkt,t}^2 + e_{it} \quad (6) \]

where

\[ r_{it} = \text{excess return of the fund} \]
\( r_{mkt,t} \) = excess return on the market benchmark
\( \lambda_{i,mkt} > 0 \) is evidence of the fund having market timing ability.

Admati, Bhattacharya, Pfleiderer, and Ross (1986) provide a theoretical foundation for this model as they derive it under the assumption that a manager has constant absolute risk aversion and changes his portfolio beta as a linear function of the signal.

In terms of the simulation, the first version of the model I test is equation (6), and I will refer to this equation as the TM model. At issue is the portfolio used to represent the market benchmark. As discussed in Chen, Ferson, and Peters (2005), there is no fixed income model equivalent to the Capital Asset Pricing Model that suggests an appropriate market portfolio. Chen et al. use style specific benchmarks to replace the market portfolio. Following this example, I use the neutral portfolio allocation (55% government/credit, 35% mortgage, and 10% cash), which was defined in section IIA, as the market portfolio.

Elton, Gruber, Das, and Hlavka (1993) demonstrate that performance evaluation results can be incorrect if the methodology does not appropriately account for the performance of assets not included in the market benchmark yet are a part of the investment opportunity set of the mutual fund manager. As a result, more recent studies (see for example, Bello and Janjigian (1997), Goetzmann, Ingersoll, and Ivkovic (2000), and Bollen and Busse (2001)) include multiple indices or factors to ensure that the manager is truly demonstrating timing skill and not exploiting weaknesses in the single index model. However, in the interest of parsimony, such studies only measure timing with respect to the market portfolio and do not include additional timing variables to measure sector or factor timing.

In terms of this simulation, the second model I examine is the two index TM model (hereafter referred to as 2TM). I include indices for both the government/credit sector and the
mortgage sector and a timing variable to measure skill relative to the government/credit sector.\textsuperscript{9} The model is as follows:

\[ r_{it} = a_i + b_i r_{gc,t} + c_i r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + e_{it} \quad (7) \]

I choose to measure timing relative to the government/credit sector due to the higher investment weight in this category as defined by the simulated portfolio strategy.

The third variation of the model is based on work done by Lehmann and Modest (1987). They demonstrate that a multi-factor extension of the TM model would include squared terms and cross terms of all the indices or factors. Grinblatt and Titman (1994), Glassman and Riddick (2004), and Comer (2005) use this approach. The models include multiple timing variables for each asset class, market, or sector that they examine. However, none of these papers examines the power of such a model.

With respect to the simulation, the multi-factor form of the TM model would be as follows:

\[ r_{it} = a_i + b_i r_{gc,t} + c_i r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + \lambda_{i,mort} r_{mort,t}^2 - d_i r_{gc,t} r_{mort,t} + e_{it} \quad (8) \]

The model includes timing variables for measuring timing skill relative to both the government/credit sector and the mortgage sector. I refer to this final version as the MFTM model. The goal of the simulation is to examine the statistical significance of $\lambda_{i,mkt}$ and $\lambda_{i,gc}$ across various managerial skill levels for each of the three models.

\section*{F. Known Empirical Problems of the Quadratic Regression Approach}

The structure of the simulation allows me to abstract from the known empirical problems with the quadratic approach and focus on the power of each of these models at identifying timing skill for various skill levels. The models are based on the fact that timing ability implies
convexity between a fund’s return and the index or factor that is used to measure timing skill. Problems with the models are due to the potential of convexities in a fund’s return series that are not related to timing ability.

Using a sample of equity funds, Ferson and Schadt (1996) illustrate that versions of the TM model are misspecified when applied to naive strategies. Conditional versions of the model that adjust for the use of public information reduce this bias in the model. However, they only focus on market timing and do not fully account for the potential of sector timing as reflected in equation (8) due to the significant amount of data required to implement their conditional approach. I do not address the issue of conditioning information in this simulation due to the data constraints. In addition, I do not model the simulated manager’s forecasts as a function of public or private information. Thus I acknowledge that no attempt is made to distinguish performance based on the use of private or public information.

A second problem with the TM approach involves the interim trading bias discussed by Goetzmann, Ingersoll, and Ivkovic (2000). Often, researchers use monthly data when estimating timing skill. GII demonstrate that if monthly data is used and the manager engages in daily timing activity, the models are weak and biased downward. Bollen and Busse (2001) use daily data to demonstrate that when a manager engages in daily timing activity tests based on higher frequency data are more powerful. In this simulation, the manager is assumed to make timing decisions on a monthly basis, and I am using monthly sector return data. Thus, the trading period matches the frequency of the return data.

A third issue is mentioned by Edelen (1999). In that study, he documents a statistically significant negative relation between a fund’s abnormal return and investor cash flows. Focusing on equity funds, he finds that cash inflows to a fund are higher when stock market returns are
high. This increases the portfolio weight on cash and causes the timing coefficients to be biased downward. The results indicate that funds experience significant negative market timing coefficients when the fund experiences cash inflows. However, in this simulation, cash flows are ignored, and it is implicitly assumed that the manager is investing the same dollar amount of assets each month across the three sectors.

A fourth concern involves the degree of convexity in the assets held by the fund relative to the convexity of the assets held in the benchmark index. Jagannathan and Korajczyk (1986) demonstrate that models detect spurious timing ability when managers hold stocks with payoffs that are more or less option like than the market proxy. I address this issue directly in the next section as I begin to examine the power of each of the models.

III. Simulation Results

In this section, I focus on two simulations. First, I examine the power of the three TM timing specifications when I assume that the manager holds the benchmark index for each of his sector allocations. I then relax the assumption and examine the size and power of the models when the manager’s sector asset selections deviate from the benchmark indices used to measure sector timing ability. These simulations allow me to examine how skill level, model specification, and choice of benchmark influence the ability to detect significant timing skill.

A. Model Power: Simulated Returns Series as Benchmark

As discussed in Section IIA, I assume the simulated fund manager has a neutral allocation of 55% government/credit, 35% mortgage, and 10% cash for each sector. If he expects a sector to be the best performing of the three, he will increase his allocation to that
sector by 10% from the neutral allocation while reducing his allocation to the sector he expects to be the worst performing. Skill level varies from 50% (neutral timing ability) to 100% (perfect timing ability) using 10% skill increments. For simplicity, if a manager has an incorrect forecast, it is assumed he increases his allocation to the worst performing sector while reducing his allocation to the best performing.

Both Jagannathan and Korajczyk (1986) and Grinblatt and Titman (1994) illustrate that performance measures are sensitive to the chosen benchmark. To ensure that the results are not influenced by the use of improper benchmark indices, I assume that the manager holds the benchmark index for each of his sector allocations. Returns for each manager \( i \) are calculated using the simulated return series as in equation (5), and the simulated return series \( r_{s gc, it} \), \( r_{s m o r t, it} \), and \( r_{s c a s h, it} \) also serve as the benchmark indices used in the model specifications.

Results for each model specification are reported in Table 3. The results illustrate that even under controlled conditions, the models are not very powerful unless the managers demonstrate exceptional skill. For each skill level, the 2TM model detects the largest percentage of significant positive timers followed by the TM model. The MFTM model is clearly the weakest. I focus on the results for the 2TM model specification given its relative power among the three models. Despite correctly forecasting bond sector returns with 70% accuracy, only 55.2% of managers are identified by the test as having significant timing skill. At the 80% level, over 30% of the managers do not generate significant timing coefficients. Once skill reaches 90%, 11.2% of managers still do not have statistically significant timing skill. The number of non significant timing coefficients at the 90% level is well above the Type I error rate of 5% and far greater than one would expect for a well specified timing test. Although not reported in the table, the percentage of significant negative timing coefficients for the 2TM model at the neutral
50% skill level is 13.2%. When compared to the percentage of positive coefficients (15.4%) for this skill level, there is no evidence that the model is biased against detecting significant positive ability.

Given that the simulation controls for skill level and benchmark accuracy, two potential reasons for the weakness of the models stand out. The lack of power may reflect that the models are poorly specified and are inappropriate for measuring sector timing skill. Alternatively, model weakness may reflect the difficulty of generating a large enough return differential relative to the benchmark index given the distribution of the sector returns.

To examine this issue, for each manager and for a specific skill level, I calculate the mean, standard deviation, and skewness of his return distribution and his benchmark. I partition the managers into two groups: significant positive timers and all others. I calculate an overall average across all managers in each partition. I then examine the difference between the average return distribution and the average benchmark return distribution of each group of managers. Because of the way the simulation is designed, I use the neutral allocation as the benchmark.

Results are reported in Table 4 for the 2TM model only but also hold for both the TM and MFTM specifications respectively. Skewness relative to the benchmark index is clearly the dominant factor influencing the likelihood of statistical significance. Across skill levels, the average skewness differential for the significant positive timers ranges from 0.303 to 0.373 indicating the convexity in returns that is detected by the model. By comparison, the average skewness differential is much smaller and only ranges from -0.024 to 0.090 for the non significant timers. The role of skewness is highlighted by a comparison across skill levels. The 155 funds that are non significant timers at the 80% skill level have a higher return differential than the 137 significant timers at the 60% skill level (0.058% vs. 0.032%). Yet, the skewness
differential of the 155 funds is smaller (0.048 vs. 0.316) and thus the funds are not identified as significant timers despite superior returns.\textsuperscript{11}

Overall, the following story emerges from these initial simulation results. The timing tests are not very powerful. Statistical significance is a function of both manager skill level and skewness of returns relative to the benchmark index. Given the typical strategies employed by managers, it is difficult for the simulated managers to generate sufficient convexity in their returns such that the tests detect significant timing skill for a large percentage of funds at reasonable skill levels.

\section*{B. Model Size and Power Revisited: Using Sector Benchmarks}

In the previous simulation, I assumed that the manager holds the benchmark index for each of his sector allocations. Very seldom will this be the case among real world managers. In this section, I repeat the simulation allowing the manager’s sector asset selections to deviate from the benchmark indices used to measure sector timing ability.

The return series for each manager are the same as in the previous simulation. However, I now use the broad sector indices (Lehman government/credit index, mortgage backed securities index, and Treasury bill index) as the benchmark indices in each of the models. Since the managers’ sector return series are based on a weighted average of the sector subindices (see equations (4) and (5)), each manager’s sector asset selections differ from those of the broad sector index.

The design of this simulation allows me to not only examine the power of the timing tests but to also examine the test size for a fixed weight monthly rebalancing strategy.\textsuperscript{12} To examine size, I assume that at the beginning of the simulation period each simulated managers holds the
neutral allocation of 55% government/credit, 35% mortgage, and 10% cash. At the end of each month, no matter what the performance of each sector, the fund manager rebalances back to the 55/35/10 mix. This monthly rebalancing strategy does not involve any active management on the part of the fund manager.

Results are presented in Table 5. The results vary greatly across the three models. The results from the TM model which uses the neutral benchmark as the market proxy are consistent with a well specified timing model. None of the 500 timing coefficients are statistically significant, thus indicating that none of the simulated managers have significant positive or perverse timing ability.

Contrary to the TM results, the other two specifications detect significant timing ability. The results from the 2TM model show some evidence of misspecification but the direction of the bias is not consistent. Significant positive ability is detected for 94 (19%) of the managers and significant negative ability for 115 (23%) of the managers. Although the detection of significant timing ability is relatively evenly distributed across positive and perverse ability, these percentages are well above the expected Type I error rate of five percent.\(^{13}\)

The results from the MFTM model show evidence of misspecification as all significant timing ability is positive. 182 (36%) of the 500 simulated managers have significant positive timing coefficients while no manager has significant negative ability. The average timing coefficient is 0.213 for the MFTM which is the only model specifications with an average positive timing coefficient. The MFTM model clearly appears to be biased toward detecting positive timing skill.

The size test results provide some initial evidence that the models may not be well specified. To examine power, I repeat the simulation conducted in Section IIIA using the broad
sector indices as the benchmarks in the models. Results are presented in Table 6 where I report the percentage of significant positive timers. For comparison, I also report the difference in the percentage of significant coefficients that was found in the previous simulation when the simulated return series served as the benchmark.

The results demonstrate that the models are weak at detecting ability for a majority of managers unless the managers demonstrate exceptional skill. According to the 2TM specification, which is the most powerful of the three models, the majority of funds do not have significant timing ability until the manager correctly forecasts sector returns 70% of the time. The TM model does not detect timing skill for the majority of managers until the 80% skill level. For all three models, more than 10% of funds do not have significant timing coefficients even at the 90% level.

The results also indicate a drop in model power from the previous simulation. The decline is consistent across all three model specifications and for virtually all skill levels. The average loss in power, as measured by the difference in the percentage of significant coefficients, is 2.5% for the 2TM model, 3.6% for the TM model, and 4.8% for the MFTM model. For the TM and 2TM models, there is a loss in power for all skill levels except the 100% skill level. Although there is a small gain in power at the 60% skill level, the MFTM model has the largest average power loss due to the model’s inability to detect timing skill for 15% of the managers who have perfect skill.

Given that all else is held constant across the simulations, the decline in model power can be attributed to the change in benchmarks. In this simulation, a fund’s sector asset selections deviate from the benchmark indices used. Thus, a fund’s characteristics can differ from those of the benchmark portfolio and affect the ability of the model to measure performance. I use the
results from the 2TM model for the 70% skill level to further explore this issue. Similar results hold for the other model specifications and skill levels.

I classify the results for each manager across the two simulations into three categories: 1) change from significant positive timing skill to non significant skill 2) no change in the significance of estimated timing skill and, 3) change from non significant timing skill to significant positive skill. Of the 500 managers, 418 (83.6%) fall into Category 2. The decline in power across simulations can be attributed to the 52 funds (10.4%) in Category 1 as opposed to only 30 funds (6.0%) falling into Category 3.

An examination of the government/credit sector subindex weights for the funds that fall into Categories 1 and 3 highlight a significant difference in the characteristics of these funds relative to the benchmark index. Funds in Category 1 have an average exposure of 59.3% to the long government/credit subindex. On average, this subindex only accounts for 27.6% of the broad government/credit index. In addition, this subindex has a skewness coefficient of -0.90 compared to a coefficient of -0.78 for the broad index. These funds have more than twice the investment weight in assets that are more negatively skewed than the benchmark index, thus accounting for the decline in model power.

The opposite occurs for the funds in Category 3. These funds have an average exposure of 51.7% to the 1-5 year government credit subindex whose returns have a skewness of -0.47. This subindex accounts for 42.6% of the broad government/credit index which has the skewness coefficient of -0.78 mentioned above. These funds have a heavier investment weight in assets that are less negatively skewed than the benchmark and are thus rewarded for timing skill when compared to the broad index.
The unequal distribution of funds between Categories 1 and 3 is due to the returns of the two subindices in which the funds are most heavily invested. The average monthly return of the government/credit index is 0.55%. The average return of the 1-5 year government credit is 0.47% while the return of the long government credit index is 0.66%. Category 3 funds benefit from the skewness of the assets, but they have a heavier investment weight in the subindex whose average returns are less than the benchmark.

The size and power results of this section highlight the potential problems with using returns based models to estimate bond sector timing skill. Overall, the results indicate that the models are relatively weak and estimated timing skill is sensitive to the benchmark used. Fund characteristics relative to the benchmark play a crucial role in the ability of the models to detect significant sector ability. As the results of the simulation indicate, detecting positive sector timing ability for bond funds is difficult with returns based models.14

IV. Attribution Return Measures Based on Portfolio Weights

The results from the previous section indicate that various specifications of the quadratic regression approach are weak and do not detect positive timing skill for sector timers until managers demonstrate a very high level of forecasting ability. As mentioned in Section II, annual investment weight by sector is available for bond funds. Of interest is whether timing measures based on portfolio weights provide a potential useful alternative to the quadratic regression approach.

In this section, I examine the power of portfolio weight based measures. Given the limited amount of available data, the goal of the weight based measures is to compare the actual return of the simulated manager to the estimated return that would have been earned had the
The manager followed a passive strategy over the time period using last period’s investment weights. The advantage to this type of measure is that it does not suffer from the empirical problems of the quadratic regression approach. But as discussed below, these measures also have their own weaknesses.

A. Attribution Return: Annual Reported Sector Weights

Various methodologies are available in the literature to examine timing ability when the researchers have access to portfolio holdings, weights, or recommendations at a high frequency (see Graham and Harvey (1996), Chance and Hemler (2001), Ferson and Khang (2002), and Jiang, Yao, and Yu (2004)). However, portfolio data related to bond funds is limited. In this study, I only have access to annual investment weights by bond market sector. Using this data, for each month $t$ during each year $y$, I estimate a monthly measure of timing skill as follows:

$$R_{\text{att},it,y} = R_{it,y} - \left( w_{gc,it,y-1} R_{gc,t} + w_{mort,it,y-1} R_{mort,t} + w_{cash,it,y-1} R_{\text{cash},t} \right)$$

I refer to $R_{\text{att},it}$ as a manager’s attribution return. The manager’s actual return during the month is compared to the return that would have been earned that month if the previous year’s reported annual portfolio sector weights were in effect. I average the monthly attribution returns and test for statistically significant differences from zero. A positive attribution return suggests that a manager is successful at sector timing.\(^{15}\)

B. Attribution Return: Average Estimated Sector Weights

The disadvantage of the previous measure is that I only have access to annual portfolio sector weights. As indicated in the simulation, a manager can engage in substantial changes to
his portfolio during the year. Thus, the annual weights may not reflect the manager’s investment policy during the year.\(^{16}\)

Sharpe (1992) suggests a methodology where a researcher can use a fund’s past return series to infer the investment policy in effect during the estimation period. Sharpe’s methodology, which is commonly referred to as returns based style analysis, represents a manager’s actual investment portfolio with the best set of asset class exposures that total to 100\% and best replicate the return series of the fund over time. I assume the monthly return for each fund can be represented by the following factor model:

\[
R_{it} = w_{gc,i} R_{gc,t} + w_{mort,i} R_{mort,t} + w_{cash,i} R_{cash,t} + e_{it} \quad (10)
\]

Then, the average exposure to each sector \(j\) during a given period can be estimated by solving the following quadratic program:

\[
\begin{align*}
\min & \quad \text{var}(R_{it} - \sum w_{ji} R_{jt}) \\
\text{subject to} & \quad 0 \leq w_{ji} \leq 1 \quad \forall \ j, i \\
& \quad \sum w_{ji} = 1
\end{align*}
\]

This technique has been applied to bond funds by Blake, Elton, and Gruber (1993) and Comer, Boney, and Kelly (2005). Both provide strong evidence that the technique can be used to infer a manager’s investment policy.

Using two year estimation windows, I use equation (11) to estimate the sector weights that most closely replicate the return series of the fund over that period. I then use these average weights in place of the actual reported weights and calculate an attribution return as in equation (9).\(^{17}\) The advantage to this approach is that I make use of the higher frequency monthly return data that is available which reflects the average allocation decisions made by the manager throughout the period. The disadvantage is that the Sharpe style weights are estimates of the
average portfolio allocation in effect over the period and are likely to estimate the true weights with some degree of error.

C. Attribution Return: Results

Using the same 500 simulated fund manager returns across the various skill levels from Section III, I examine the ability of both attribution return measures to detect significant positive timing ability. I compute the attribution return measure comparing actual fund returns against a fixed weight monthly rebalancing strategy with weights changing on an annual basis. To calculate the monthly attribution return series for 1998, I use the neutral allocation weights to represent the lagged weights since the simulated data begins in 1998.

Results are reported in Table 7. For each specification, I report the average annualized attribution return across all 500 simulations and the number of positive and statistically significant attribution returns. Only one manager at the 50% skill level has a statistically significant negative attribution return. Thus, I do not report the number of simulated managers with negative timing skill.

The results are consistent with results one would expect from a well specified timing methodology and appear much more powerful than the results from the quadratic regression approach. Both attribution return measures have substantial power. Overall, the style weight measure appears to be the most powerful despite having a lower average attribution return for each skill level. At the 50% skill level, both models signal that only a small percentage of managers (0.4% for the style weight measure and 3.8% for the portfolio weight measure) have significant positive timing skill. At the 70% skill level, both models signal that over 69% of
managers have timing skill which is in direct contrast to the results from the TM model specifications.

Although both models have considerable power, the difference in the two measures is pronounced beginning at the 70% skill level. Once skill reaches 70%, the style weight measure indicated that 92.2% of managers have significant timing ability. At the 80% through perfect skill levels, the style weight measure indicates that all managers have timing skill. The percentages of significant timers are smaller for the portfolio weight measure. The measure identifies 85%, 90%, and 93% of managers as significant timers for the 80%, 90%, and 100% skill levels respectively.

These results suggest that an attribution return measure provides a more accurate measure of sector timing skill among bond funds than models based on the quadratic regression approach. Given the problems with the quadratic approach that were discussed in Section IIF, the attribution measure may be a better approach despite the limited data available on bond portfolio sector holdings.

V. Weight Based Measures and Actual Fund Performance

In this final section, I turn to using the attribution measure to measure the skill of the sample of general government bond funds whose portfolio data provided the basis for the simulation. I compare actual fund performance to two passive investment strategies: 1) monthly rebalancing with weights reset annually, and 2) buy and hold with weights reset annually. I use both annual reported sector weights and estimated average biannual sector style weights. In addition, I use the estimated gross fund return that was first calculated in Section IIE.
Results are reported in Table 8. As constructed, the sample suffers from survivorship bias as the funds were selected from those that were in existence over the entire 1998-2004 time period. However, the degree of potential survivorship bias is not severe. According to Morningstar data, only eight general government bond funds were in existence as of December 31, 1997 that did not survive to the end of 2004. It is unlikely that the conclusions drawn would change even if the portfolio data were available to include these funds in the analysis.

I first focus on the portfolio weight attribution return measure. The results are virtually identical across both passive investment strategies and indicate no evidence of significant positive or perverse timing ability. Across all 122 funds, the average annualized attribution return is -0.0009% for the monthly rebalancing strategy and -0.001% for the buy and hold strategy. According to the buy and hold strategy, six funds have significant positive attribution returns compared to three funds with significant negative returns. The numbers are four and three respectively for the buy and hold strategy. The distribution of positive and negative attribution returns is the same (59 positive and 63 negative) across both passive strategies.

Because I am examining the skill of actual funds, I need to use an alternative index model to estimate the biannual style weights for the style based attribution return measure. I use the following eight index model

\[
R_{it} = w_{\text{cash},i}R_{\text{cash},t} + w_{\text{gc0105},i}R_{\text{gc0105},t} + w_{\text{gc0510},i}R_{\text{gc0510},t} + w_{\text{gelong},i}R_{\text{gelong},t} + w_{\text{hiyld},i}R_{\text{hiyld},t} \\
+ w_{\text{mort},i}R_{\text{mort},t} + w_{\text{globe},i}R_{\text{globe},t} + e_i \quad (12)
\]

where

- \( R_{\text{cash}} \) = total return on the Treasury bill index
- \( R_{\text{gc0105}} \) = total return on the 1-5 year government/credit bond index
- \( R_{\text{gc0510}} \) = total return on the 5-10 year government/credit bond index
- \( R_{\text{gelong}} \) = total return on the long (10+) government/credit bond index
This model addresses potential weaknesses of the three index model. The model is based on the seven index model used by Comer, Boney, and Kelly (2005) and includes an additional term to capture the impact of foreign bond investment. As discussed in Section IIIB, the sample of funds may hold government/credit bonds of various maturities that behave differently than the broader government/credit index used in the three index model. Second, Morningstar does not report the portfolio weight allocated to high yield bonds. According to fund prospectuses, these funds are allowed to invest in this sector and potentially may have a substantial portion of assets allocated to this sector. The model is more likely to provide estimated weights that reflect the true portfolio strategy employed by the individual fund.

Results are also presented in Table 8. Again, results are similar for both the monthly rebalancing and buy and hold strategies, and I discuss the monthly rebalancing results below. The simulation results suggest that the style measure is more powerful, and the results indicate that the sample may have some positive timing skill. The average annualized attribution return is 0.0018%. Eleven (9.0%) of the funds have significant positive attribution returns while no fund has a significant negative return. A comparison across the two measures indicates that five of the six funds with significant positive attribution returns according to the portfolio weight measure continue to have significant attribution returns under the style measure.

The general result across these two attribution return measures is that at worse managers have neutral skill and may have some positive sector timing ability. The results are also consistent with the result from Section IID where I found that the return distribution of the fund
sample most closely mimics the simulated returns series from the neutral skill level. Reconciling the differences between the two measures is difficult without further detailed information concerning the fund’s portfolios. But there appears to be no evidence suggesting that the fund sample has perverse sector timing skill.

VI. Conclusion

The returns based model of Treynor and Mazuy (1966) was designed to examine a manager’s ability to switch assets between a broad market portfolio of stocks and Treasury bills. Limited portfolio data available from bond funds indicates that managers engage in sector timing behavior. This paper focuses on the ability of this model and its multi index extensions to measure timing ability when managers explicitly engage in sector timing behavior.

I generate simulated fund returns for 500 managers using portfolio strategies based on annual reported sector allocations. The size and power of the methodology are examined along three dimensions: 1) managerial skill level, 2) the benchmark used to measure performance, and 3) the version of the quadratic model used to estimate skill. Results indicate that the various versions of the TM model are weak, and managers must have substantial skill (80% and higher) before the model is likely to signal significant positive timing ability for the majority of funds. Given the typical strategies employed by managers and the return distribution of the bond sectors, it is difficult for the simulated managers to generate sufficient convexity in their returns such that the tests will detect significant timing skill for a large percentage of funds at reasonable skill levels.

An alternative measure that I refer to as an attribution return, which compares actual monthly fund performance to a passive strategy based on annual or estimated average sector
weights is better able to detect positive ability for a majority of funds at reasonable skill levels
despite the limited number of portfolio observations available. When applied to a sample of
general government bond funds, the attribution return measure indicates that at worse managers
have neutral skill and may have some positive timing ability.

These results indicate that researchers should exert caution in drawing conclusions
concerning portfolio performance when returns based models are used to evaluate bond fund
managers who explicitly engage in sector timing behavior. The results also add to the growing
literature that suggests portfolio based models have superior ability to detect positive timing skill
when it exists.
This importance is highlighted by data from the Investment Company Institute. From the end of 1997 to the end of 2004 (the time period to be covered in this study) assets under management in domestic nonmunicipal bond funds have grown by nearly 120% from $426.3 billion to $925.8 billion.

These papers look at different aspects of timing than the sector timing behavior examined in this study. Chen et al. look at timing skill with respect to factors related to bond returns. Comer et al. examine the ability of managers to correctly adjust the average maturity of the portfolio.

I choose to focus on the Treynor Mazuy model and not the Henriksson Merton model. The academic literature has provided a theoretical justification for a multi-index extension of the TM model (see Lehmann and Modest (1987)) and various versions of this extension have appeared in the literature. The equivalent can not be said for the HM model.

The largest Morningstar category is general corporate bond funds. Unfortunately, due to the voluntary nature of the portfolio reporting, the data for this sector is less complete than the data for general government funds. The complete set of annual weights is missing for nearly 10% (compared to 5.4% for the government funds) of the potential corporate sample. However, the data for the available funds does indicate substantial sector timing activity. The average range of reported sector weights varies from 17.8% for the cash sector to 25.3% for the government/corporate sector which compares favorably with the data reported in Table 1 for the government sector.

The entire list of general government bond funds includes 427 funds. Once exchange traded funds and multiple classes of the same fund are excluded, there are 148 funds with distinct portfolios. 129 funds have been in existence over the entire time period.

I note that most market timing studies examine equity fund performance over time periods of at least 10 years. The lengthy observation periods are typically motivated more by empirical and statistical issues rather than the average tenure of a specific fund manager. Thus, these studies are examining the performance of a fund more than a specific manager. The goal of this study is to reflect the typical environment of bond fund managers and then examine the power of return based models to measure timing.

I formally test that the returns to each of the sectors is normally distributed. I calculate the Jarque Berra statistic which is distributed chi squared under the null. The value of the test statistic varies from 40.6 for the government/credit sector to 76.3 for the mortgage sector. A value of 9.2 or higher rejects the null hypothesis of normality.

I note that this return is still an estimate. It is likely to underestimate the true gross fund return because I have not adjusted for the costs associated with fund turnover. I will use this same gross fund return later in Section V when I examine the performance of the fund sample.

As mentioned earlier in the section, there is no established fixed income equivalent to the CAPM or the four factor model of Carhart (1997). Given the way that the simulation is designed, the most appropriate extension of the timing models would involve the use of indices rather than factors since the simulated managers (and actual managers given the reported portfolio holdings) are known to time across the specific sectors.

Ferson and Schadt (1996) examine timing performance from 1968-1990 which provides a total of 276 monthly observations. Their conditional version of the TM model includes the product of lagged public information variables with the excess market return as additional explanatory variables in the regression model. They include five public information variables in the model. Given this study is based on 84 monthly observations, parsimony is an important consideration in the selection of models to evaluate.

I more formally test this using a logit model. The dependent variable equals 1 if the fund’s timing coefficient is statistically significant and equals 0 otherwise. The independent variables are 1) the difference between fund return and benchmark return, 2) the difference between fund return skewness and benchmark skewness, and 3) dummy variables for each skill level. The coefficient for the skewness variable and each of the skill level dummy variables are statistically significant and the pseudo r-squared of the model is 0.91.

In the previous simulation, the asset selections of the managers are identical to the benchmark index. Thus, by definition, a fixed weight monthly rebalancing strategy would produce a perfect fit among the timing regressions.

The results for the 2TM model are consistent with results for a set of synthetic equity funds examined by Bollen and Busse (2001) when using an equivalent model specification. Using daily data, they construct synthetic equity funds that have no timing ability by construction. They find that over 45% of their synthetic sample has statistically significant timing ability when they use a four factor version of the TM model with a timing variable for only the stock market portfolio.
As a robustness check, I substituted data from the high yield sector for the mortgage sector and repeated the simulation results. The primary difference between the high yield and mortgage sector is their correlation with the government/credit sector. The correlation between high yield and government/credit is only .08 and high yield has a negative correlation with the cash sector. For the sake of brevity the results are not reported here. But all simulation results were qualitatively similar. Thus, the results are not driven by the sectors used in this study.

I acknowledge that positive attribution returns can be generated from two sources: 1) strategic changes to the portfolio sector allocations (sector timing ability) or 2) sector bond selections that deviate from the benchmark indices included in the model (asset selection ability). The simulation is designed such that the asset selection ability of a manager is fixed over the time period since the investment weights for each sub sector index are held constant over the entire sample period. Thus, the attribution return measure reflects actual sector timing skill more than asset selection ability.

Morey and O’Neal (2004) provide some evidence that bond fund managers may engage in window dressing ahead of portfolio disclosure dates.

For example, I use the simulated monthly return series covering 1998-1999 and equation (11) to estimate the average sector weights in effect over the period. These weights are then used in place of the actual reported year end 1999 weights to calculate the monthly 2000 attribution returns.

I also compare actual fund returns to three other passive strategies: 1) buy and hold with annual rebalancing to the reported weights, 2) monthly rebalancing where the initial portfolio weights are fixed forever, and 3) buy and hold for the entire time period using the initial weights. The results are virtually identical across strategies and for the sake of brevity are not reported here.

I compute the average annualized attribution return as follows. I first calculate the average monthly attribution return for each of the 500 simulated managers. I then average across all 500 managers to obtain an average monthly attribution return. Finally, I multiply this value by 12 to obtain an annualized average.

I do not measure skill of the actual fund sample using the TM model. As discussed in Section IIF, several adjustments to the TM methodology must be made when dealing with actual fund returns. These adjustments are beyond the scope of this paper. The reader is referred to Chen, Ferson, and Peters (2005).

In order to compute the monthly 1998 attribution returns, I need to estimate the style weights using return data for 1996-1997. Because the sample was based on funds in existence as of 1997, four funds in the sample do not have return data available for 1996. For these funds, I used the reported end of 1997 weights to compute the 1998 attribution returns. Results are not affected if these funds are excluded from the sample.
This data provides descriptive statistics of the reported sector investment weights of the sample of general government bond funds. The investment weights are reported by Morningstar and are based on voluntary portfolio data provided by the funds to Morningstar. The sample is composed of the 122 funds that have a history of annual investment weights available over the 1998-2004 time period. To obtain the values in the table, I first use the time series of weights to calculate the statistic for each individual fund and I then average each statistic across the entire sample of funds. Range is defined as the difference between the maximum and minimum reported weights for each specific category. The sum of the mean and median investments weights does not equal 100% due to the exclusion of the foreign sector which has a mean weight of 1.0%

<table>
<thead>
<tr>
<th>Sector</th>
<th>Mean weight</th>
<th>Median weight</th>
<th>Standard Dev</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Govt/credit</td>
<td>53.7%</td>
<td>54.6%</td>
<td>12.3%</td>
<td>33.4%</td>
</tr>
<tr>
<td>Mortgage</td>
<td>37.5%</td>
<td>37.3%</td>
<td>10.7%</td>
<td>29.1%</td>
</tr>
<tr>
<td>Cash</td>
<td>7.8%</td>
<td>6.8%</td>
<td>7.3%</td>
<td>19.1%</td>
</tr>
</tbody>
</table>
Table 2
Comparison of Distribution of Actual and Simulated Monthly Fund Returns

The table compares the distribution of the monthly returns of the actual fund sample to monthly simulated returns for three managerial skill levels over the 1998-2004 sample period. The statistics are calculated for each individual fund and then averaged across the funds to obtain the values reported in the table. The statistics for the fund sample are based on a sample of 122 general government bond funds taken from the March 2005 Morningstar Principia CD. Monthly fund returns are from CRSP and are reported as percentage returns. The monthly fund returns are estimate gross returns as estimated fund expenses have been added back to the reported fund return. Statistics for simulated returns are reported for three skill levels represented by 1) perverse timing ability (0% skill), 2) neutral timing ability (50% skill), and 3) perfect timing ability (100% skill). Each simulated manager follows a strategy where he is assumed to have a neutral allocation for each sector of 55% government/credit, 35% mortgage, and 10% cash. If he expects a sector to be the best performing of the three, he increases his allocation to that sector by 10% from the neutral allocation and reduces the allocation to the worst sector. An incorrect forecast is defined as the manager predicting the reverse order of the actual relative performance of the three sectors. The simulated fund return for manager \(i\) for month \(t\) is given by

\[ R_{it} = w_{gc,i}R_{sgc,it} + w_{mort,i}R_{smort,it} + w_{cash,i}R_{scash,it} \]

where \(w_{gc,i}\), \(w_{mort,i}\), and \(w_{cash,i}\) are the weights based on the portfolio strategy and skill levels in effect and \(R_{sgc,it}\), \(R_{smort,it}\), and \(R_{scash,it}\) reflect the returns from the simulated sector portfolios held by each manager as defined in the text.

<table>
<thead>
<tr>
<th>Fund sample</th>
<th>Perverse skill</th>
<th>Neutral skill</th>
<th>Perfect skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.510</td>
<td>0.415</td>
<td>0.531</td>
</tr>
<tr>
<td>Standard dev</td>
<td>0.98</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.571</td>
<td>-1.299</td>
<td>-0.875</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1.260</td>
<td>3.081</td>
<td>2.114</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.57</td>
<td>2.24</td>
<td>2.71</td>
</tr>
<tr>
<td>Minimum</td>
<td>-2.88</td>
<td>-3.87</td>
<td>-3.38</td>
</tr>
</tbody>
</table>
The table reports results for the 500 TM timing tests for each skill level and each timing specification when the simulated manager employs an active allocation strategy and each manager’s simulated return series serve as the benchmark indices used in the model specifications. Simulated fund returns for manager $i$ during month $t$ are generated by

$$R_{it} = w_{gc, it} R_{sgc, it} + w_{mort, it} R_{smort, it} + w_{cash, it} R_{scash, it}$$

where $w_{gc, it}$, $w_{mort, it}$, and $w_{cash, it}$ are the weights based on the portfolio strategy and skill levels in effect and $R_{sgc, it}$, $R_{smort, it}$, and $R_{scash, it}$ reflect the returns from the simulated sector portfolios held by each manager as defined in the text. $R_{sgc, it}$, $R_{smort, it}$, and $R_{scash, it}$ also serve as the benchmark indices used in the model specifications. Each simulated fund manager is assumed to have a neutral allocation of 55% government/credit, 35% mortgage, and 10% cash for each sector. If he expects a sector to be the best performing of the three, he increases his allocation to that sector by 10% from the neutral allocation while reducing his allocation to the sector he expects to be the worst performing. If a manager has an incorrect forecast, it is assumed he increases his allocation to the worst performing sector while reducing his allocation to the best performing sector. Skill level varies from 50% (neutral timing ability) to 100% (perfect timing ability) using 10% skill increments. Individual fund timing coefficients are estimated for each of the simulated managers using each of the following TM specifications where $r_{jt}$ represents excess returns:

$$r_{jt} = a_i + b_i r_{mkt, t} + \lambda_{i,mkt} r_{mkt, t}^2 + e_{it}$$  \hspace{1cm} (6)$$

$$r_{jt} = a_i + b_i r_{gc, t} + c_i r_{mort, t} + \lambda_{i,gc} r_{gc, t}^2 + e_{it}$$  \hspace{1cm} (7)$$

$$r_{jt} = a_i + b_i r_{gc, t} + c_i r_{mort, t} + \lambda_{i,gc} r_{gc, t}^2 + \lambda_{i,mort} r_{mort, t}^2 - d_i r_{gc, t} r_{mort, t} + e_{it}$$  \hspace{1cm} (8)$$

which represent the TM, 2 index TM, and multifactor TM timing models respectively. $r_{mkt, t}$ reflects the excess return to the neutral allocation. The average timing coefficient across all 500 simulations is reported. Statistical significance is based on the quadratic term used to measure timing ability ($\lambda_{i,mkt}$ for the TM model and $\lambda_{i,gc}$ for the 2TM and MFTM models). Significance of the timing variables is based on heteroskedastic consistent standard errors and reflect the five percent level of significance.

<table>
<thead>
<tr>
<th>Skill Level</th>
<th>TM Model</th>
<th>2TM Model</th>
<th>MFTM Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Coefficient</td>
<td>Percent Sig Positive</td>
<td>Average Coefficient</td>
</tr>
<tr>
<td>50%</td>
<td>0.160</td>
<td>13.6%</td>
<td>0.090</td>
</tr>
<tr>
<td>60%</td>
<td>0.973</td>
<td>24.6%</td>
<td>0.477</td>
</tr>
<tr>
<td>70%</td>
<td>2.439</td>
<td>51.4%</td>
<td>1.229</td>
</tr>
<tr>
<td>80%</td>
<td>3.060</td>
<td>66.7%</td>
<td>1.519</td>
</tr>
<tr>
<td>90%</td>
<td>4.196</td>
<td>87.4%</td>
<td>2.074</td>
</tr>
<tr>
<td>100%</td>
<td>5.015</td>
<td>100.0%</td>
<td>2.475</td>
</tr>
</tbody>
</table>
Table 4
Simulated Manager and Benchmark Return Distributions: A Comparison of Significant and Nonsignificant Timers

The table reports results highlighting the difference in the average return distribution for significant positive timers. Results are reported for the 2TM model. For each skill level, the simulated managers are partitioned into two groups: significant positive timers and all others as identified by the simulation results. For each manager within a partition, I calculate the difference in descriptive statistics (mean, standard deviation, skewness) between the manager’s return distribution and the return distribution of the benchmark. The neutral allocation (55% government/credit, 35% mortgage, and 10% cash) serves as the benchmark.

<table>
<thead>
<tr>
<th>50% Skill Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Obs</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Skewness</td>
</tr>
<tr>
<td>Sig Positive</td>
<td>77</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.303</td>
</tr>
<tr>
<td>All Others</td>
<td>423</td>
<td>-0.001</td>
<td>0.014</td>
<td>-0.024</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>60% Skill Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Obs</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Skewness</td>
</tr>
<tr>
<td>Sig Positive</td>
<td>137</td>
<td>0.032</td>
<td>-0.012</td>
<td>0.316</td>
</tr>
<tr>
<td>All Others</td>
<td>363</td>
<td>0.018</td>
<td>0.017</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>70% Skill Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Obs</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Skewness</td>
</tr>
<tr>
<td>Sig Positive</td>
<td>276</td>
<td>0.054</td>
<td>0.001</td>
<td>0.332</td>
</tr>
<tr>
<td>All Others</td>
<td>224</td>
<td>0.041</td>
<td>0.019</td>
<td>0.046</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>80% Skill Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Obs</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Skewness</td>
</tr>
<tr>
<td>Sig Positive</td>
<td>345</td>
<td>0.074</td>
<td>-0.004</td>
<td>0.342</td>
</tr>
<tr>
<td>All Others</td>
<td>155</td>
<td>0.058</td>
<td>0.025</td>
<td>0.048</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>90% Skill Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of Obs</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Skewness</td>
</tr>
<tr>
<td>Sig Positive</td>
<td>444</td>
<td>0.096</td>
<td>-0.004</td>
<td>0.373</td>
</tr>
<tr>
<td>All Others</td>
<td>56</td>
<td>0.078</td>
<td>0.002</td>
<td>0.090</td>
</tr>
</tbody>
</table>
Table 5
Simulation Results: Size of TM Timing Tests Using Sector Benchmarks

The table reports results for the 500 TM timing tests for various manager strategies. Simulated fund returns for manager \(i\) during month \(t\) are generated by

\[
R_{it} = w_{gc,it} R_{sgc,it} + w_{mort,it} R_{smort,it} + w_{cash,it} R_{scash,it}
\]

where \(w_{gc,it}\), \(w_{mort,it}\), and \(w_{cash,it}\) are the sector weights and \(R_{sgc,it}\), \(R_{smort,it}\), and \(R_{scash,it}\) reflect the returns from the simulated sector portfolios held by each manager as defined in the text. Each simulated manager is assumed to employ a passive strategy where he holds a neutral allocation of 55% government/credit, 35% mortgage, and 10% cash. These weights are held fixed for the entire sample period (i.e. a monthly rebalancing strategy). Individual fund timing coefficients are estimated for each of the simulated managers using each of the following TM specifications where \(r_j\) represents excess returns:

\[
\begin{align*}
\lambda_{it} &= \alpha_i + \beta_i r_{mkt,t} + \lambda_{i,mkt} r_{mkt,t}^2 + \epsilon_{it} \\
\lambda_{it} &= \alpha_i + \beta_i r_{gc,t} + \gamma_i r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + \epsilon_{it} \\
\lambda_{it} &= \alpha_i + \beta_i r_{gc,t} + \gamma_i r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + \lambda_{i,mort} r_{mort,t}^2 - \delta_i r_{gc,t} r_{mort,t} + \epsilon_{it}
\end{align*}
\]

which represent the TM, 2 index TM, and multifactor TM timing models respectively. The Lehman Brothers government/credit index, mortgage backed securities index, and Treasury bill index are used as the benchmarks returns in the model. The mean coefficient reported below is the average of the individual coefficients on the quadratic term used to measure timing ability (\(\lambda_{i,mkt}\) for the TM model and \(\lambda_{i,gc}\) for the 2TM and MFTM models). Statistical significance of the timing variables is based on heteroskedastic consistent standard errors and reflect the five percent level of significance.

<table>
<thead>
<tr>
<th>Panel A: Size Tests</th>
<th>Mean timing coefficient</th>
<th>Percentage Significant</th>
<th>Significant Positive</th>
<th>Significant Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>-0.340</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2TM</td>
<td>-0.198</td>
<td>41.8%</td>
<td>18.8%</td>
<td>23.0%</td>
</tr>
<tr>
<td>MFTM</td>
<td>0.213</td>
<td>36.4%</td>
<td>36.4%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Table 6
Simulation Results: Power of TM Timing Tests Using Sector Benchmarks

The table reports results for the 500 TM timing tests for each skill level and each timing specification when the simulated manager employs an active allocation strategy. Simulated fund returns for manager $i$ during month $t$ are generated by

$$R_{it} = w_{gc,it}R_{gc,it} + w_{mort,it}R_{mort,it} + w_{cash,it}R_{cash,it}$$

where $w_{gc,it}$, $w_{mort,it}$, and $w_{cash,it}$ are the weights based on the portfolio strategy and skill levels in effect and $R_{gc,it}$, $R_{mort,it}$, and $R_{cash,it}$ reflect the returns from the simulated sector portfolios held by each manager as defined in the text. Each simulated fund manager is assumed to have a neutral allocation of 55% government/credit, 35% mortgage, and 10% cash for each sector. If he expects a sector to be the best performing of the three, he increases his allocation to that sector by 10% from the neutral allocation while reducing his allocation to the sector he expects to be the worst performing. If a manager has an incorrect forecast, it is assumed he increases his allocation to the worst performing sector while reducing his allocation to the best performing sector. Skill level varies from 50% (neutral timing ability) to 100% (perfect timing ability) using 10% skill increments. Individual fund timing coefficients are estimated for each of the simulated managers using each of the following TM specifications where $r_{jt}$ represents excess returns:

$$r_{it} = a_i + b_j r_{mkt,t} + \lambda_{i,mkt} r_{mkt,t}^2 + e_{it}$$ (6)

$$r_{it} = a_i + b_j r_{gc,t} + c_j r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + e_{it}$$ (7)

$$r_{it} = a_i + b_j r_{gc,t} + c_j r_{mort,t} + \lambda_{i,gc} r_{gc,t}^2 + \lambda_{i,mort} r_{mort,t}^2 - d_j r_{gc,t} r_{mort,t} + e_{it}$$ (8)

which represent the TM, 2 index TM, and multifactor TM timing models respectively. $r_{mkt,t}$ reflects the excess return to the neutral allocation. The percentage of significant positive timing coefficients across all 500 simulations is reported. Statistical significance is based on the quadratic term used to measure timing ability ($\lambda_{i,mkt}$ for the TM model and $\lambda_{i,ge}$ for the 2TM and MFTM models). Significance of the timing variables is based on heteroskedastic consistent standard errors and reflect the five percent level of significance. The Lehman Brothers government/credit index, mortgage backed securities index, and Treasury bill index are used as the benchmarks in the model. For comparison, the difference in percentage of significant coefficients from the previous simulation where the simulated returns series was used as the benchmark is also reported.

<table>
<thead>
<tr>
<th>Skill Level</th>
<th>TM Model</th>
<th>2TM Model</th>
<th>MFTM Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent Sig Positive</td>
<td>Difference from Sim Benchmark</td>
<td>Percent Sig Positive</td>
</tr>
<tr>
<td>50%</td>
<td>11.0%</td>
<td>-2.6%</td>
<td>13.2%</td>
</tr>
<tr>
<td>60%</td>
<td>20.6%</td>
<td>-4.0%</td>
<td>23.4%</td>
</tr>
<tr>
<td>70%</td>
<td>45.6%</td>
<td>-5.8%</td>
<td>50.8%</td>
</tr>
<tr>
<td>80%</td>
<td>60.2%</td>
<td>-6.5%</td>
<td>65.4%</td>
</tr>
<tr>
<td>90%</td>
<td>85.0%</td>
<td>-2.4%</td>
<td>88.4%</td>
</tr>
<tr>
<td>100%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

41
Table 7
Simulation Results: Attribution Return Measure and Simulated Allocation Strategies

The table reports results for the 500 attribution return timing tests for each skill level when the simulated manager employs an active allocation strategy. The simulated fund manager is assumed to have a neutral allocation of 55% government/credit, 35% mortgage, and 10% cash for each sector. If he expects a sector to be the best performing of the three, he will increase his allocation to that sector by 10% from the neutral allocation while reducing his allocation to the sector he expects to be the worst performing. If a manager has an incorrect forecast, it is assumed he increases his allocation to the worst performing sector while reducing his allocation to the best performing sector.

Skill level varies from 50% (neutral timing ability) to 100% (perfect timing ability) using 10% skill increments. The monthly portfolio weight attribution return measure is calculated for each month $t$ during each year $y$ as:

$$ R_{att, t, y} = R_{it, y} - \left( w_{gc, it, y-1} R_{gc, t} + w_{mort, it, y-1} R_{mort, t} + w_{cash, it, y-1} R_{cash, t} \right) $$

where $w_{gc, it}$, $w_{mort, it}$, and $w_{cash, it}$ are the annual sector portfolio weights as reported to Morningstar. The style weight measure is calculated using equation (9) except estimated biennial weights replace the reported weights. The biennial weights are estimated using Sharpe’s style analysis as described in the text. The average attribution return reported in table is the average across the 500 funds and has been annualized. Statistical significance is based on a standard t-test and uses the five percent level of significance.

<table>
<thead>
<tr>
<th>Skill</th>
<th>Portfolio Weight Measure</th>
<th>Style Weight Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Att Return</td>
<td>Percentage Sig Pos</td>
</tr>
<tr>
<td>50%</td>
<td>0.22</td>
<td>3.8</td>
</tr>
<tr>
<td>60%</td>
<td>0.48</td>
<td>27.8</td>
</tr>
<tr>
<td>70%</td>
<td>0.78</td>
<td>69.4</td>
</tr>
<tr>
<td>80%</td>
<td>1.04</td>
<td>85.2</td>
</tr>
<tr>
<td>90%</td>
<td>1.32</td>
<td>90.4</td>
</tr>
<tr>
<td>100%</td>
<td>1.57</td>
<td>93.2</td>
</tr>
</tbody>
</table>
Table 8
Attribution Return Measure and Fund Performance

The table reports results for the attribution return timing tests for the sample of 122 general government bond funds. The monthly portfolio weight attribution return measure is calculated for each fund for each month \( t \) during each year \( y \) as:

\[
R_{\text{att},it,y} = R_{it,y} - \left( w_{\text{gc},it,y-1} R_{\text{gc},it} + w_{\text{mort},it,y-1} R_{\text{mort},it} + w_{\text{cash},it,y-1} R_{\text{cash},it} \right)
\]

where \( w_{\text{gc},it} \), \( w_{\text{mort},it} \), and \( w_{\text{cash},it} \) are the annual sector portfolio weights as reported to Morningstar. The eight sector measure is based on an index model that includes the three government/credit maturity indices in place of the broad government/credit index, a high yield index, and a global index. Estimated biannual sector weights replace the reported weights in the calculation of the attribution return. The biennial weights are estimated using Sharpe’s style analysis as described in the text. The average attribution return reported in table is the average across the 122 funds and has been annualized. Statistical significance is based on a standard t-test and uses the five percent level of significance.

<table>
<thead>
<tr>
<th>Panel A: Monthly Rebalancing Strategy</th>
<th>Portfolio Weight</th>
<th>Eight Sector Style</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized att ret</td>
<td>-0.00092</td>
<td>0.00108</td>
</tr>
<tr>
<td>% positive</td>
<td>48.4</td>
<td>59.0</td>
</tr>
<tr>
<td>% significant positive</td>
<td>4.9</td>
<td>9.0</td>
</tr>
<tr>
<td>% negative</td>
<td>51.6</td>
<td>41.0</td>
</tr>
<tr>
<td>% significant negative</td>
<td>2.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Buy and Hold Strategy</th>
<th>Portfolio Weight</th>
<th>Eight Sector Style</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized att ret</td>
<td>-0.00105</td>
<td>0.00126</td>
</tr>
<tr>
<td>% positive</td>
<td>48.4</td>
<td>59.0</td>
</tr>
<tr>
<td>% significant positive</td>
<td>3.3</td>
<td>9.0</td>
</tr>
<tr>
<td>% negative</td>
<td>51.6</td>
<td>41.0</td>
</tr>
<tr>
<td>% significant negative</td>
<td>2.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>
References


