Available online at www.sciencedirect.com
SCIENCE DIRECT
North American Journal of Economics and Finance
17 (2006) 207–230
THE NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE

Performance evaluation with portfolio holdings information

Russ Wermers *

Department of Finance, Robert H. Smith School of Business, University of Maryland at College Park, College Park, MD 20742-1815, USA

Received 22 November 2004; received in revised form 15 January 2006; accepted 27 January 2006
Available online 23 February 2006

Abstract

This paper surveys recent academic research that uses portfolio holdings to evaluate the performance of an asset manager. These approaches mitigate the benchmark-choice problem of Roll (1978), as well as providing a much more precise attribution of the sources of manager returns. Although originally developed with U.S. data, recent papers have applied these approaches to European, Asian, and Australian equity managers. All surveyed approaches can be integrated into the Brinson, Hood, and Beebower (1986) attribution method, if we allow the composition of the benchmark portfolio to evolve through time according to the observed portfolio holdings of an asset manager.

JEL classification: G11

Keywords: Mutual funds; Performance evaluation; Investment management

1. Introduction

U.S. institutional investors (mutual funds, pension funds, endowments, trusts, and insurance companies) currently control over $10 trillion in stocks. So-called active fund managers control the great majority of these stock portfolios (roughly 90% of total assets of mutual funds in year 2000)—these managers aggressively add stocks to (and liquidate stocks from) their portfolios in an attempt to beat the S&P 500 index. Their fund families just as aggressively advertise the track records of well-performing funds in an attempt to attract new money from retail investors.

* Tel.: +1 301 405 0572; fax: +1 301 405 0359.
E-mail address: rwermers@rhsmith.umd.edu.

1062-9408/$ – see front matter
doi:10.1016/j.najef.2006.01.001

For example, recent television ads in the U.S. for the Fidelity family of funds prominently feature the well-known, former Fidelity Magellan fund manager, Peter Lynch.

The other side of the active-management story is told most aptly (and, perhaps most frequently!) by John Bogle, former Chairman of the Vanguard fund family. Mr. Bogle argues the virtues of index funds, and states that actively managed funds waste investor money by trading too frequently and by charging excessive expenses. This message has not been lost on investors. The Vanguard Index 500 fund is now one of the largest funds in the U.S., with roughly \$110 billion under management in 2006, greater than that of the Magellan fund (\$52 billion).

Academic financial economists have also been keenly interested in the value of active portfolio management at least since the seminal paper of Jensen (1968). Recently, this issue has seemingly become settled, since almost all studies that use returns-based performance analysis methods conclude that actively managed mutual funds (e.g., the Fidelity Magellan fund), on average, underperform their passively managed counterparts (e.g., the Vanguard Index 500 fund). For example, the widely cited Carhart (1997) study finds that the average actively managed U.S. mutual fund underperforms its benchmarks by over 1% on a net return basis.[1] However, many individual investors and institutional sponsors are concerned simply with evaluating the set of managers that they currently employ, or might employ in the future to manage their assets. In turn, the past underperformance of the average actively managed fund does not preclude finding subgroups of managers with skills.[2] In addition, returns-based methods used in the past, while perhaps adequately addressing the overall performance of the entire active-management industry, or large subgroups of funds, may be too noisy to work well with individual funds or small groups of funds, especially those with a limited history of returns.

Recent advances in methods that examine performance at the security level allow researchers (and investors) to paint a much more detailed picture of the performance of asset managers; these advances make a re-examination of the active/passive issue possible in much the same way that advances in DNA-profiling re-open many criminal cases for a more thorough analysis of the probable guilt of a defendant beyond more traditional fingerprinting methods. These security-level performance evaluation approaches have become known as portfolio holdings-based performance evaluation and attribution or simply holdings-based performance measurement.

There are several reasons why the use of portfolio holdings data may provide new insights into managed performance, relative to returns-based methods. First, contrary to returns-based methods, recent approaches that use portfolio holdings data address the benchmark-error problem outlined by Roll (1978). Specifically, in a mean-variance setting, Roll demonstrates how a benchmark formed from any fixed set of assets can be chosen with similar properties (but different weights on component assets) to a second benchmark, such that the ranking (by the regression alpha) of a group of managers is reversed when moving from the first to the second benchmark, as long as the two benchmarks are both inefficient. Since a truly efficient benchmark is impossible to locate with perfect precision, any ranking of managers with real-world proxy portfolios chosen to represent this unknown efficient portfolio will be flawed; therefore, we cannot blindly trust performance evaluation using market proxies for the true theoretical benchmarks. Returns-based methods cannot overcome this benchmark-choice problem, while portfolio holdings-based methods can.

Second, the style-orientation of a fund may shift non-trivially during short time periods—such as a value manager who adds technology stocks during the heyday of growth stocks of the 1990s.

[1] Other widely cited studies that examine net returns include Brown and Goetzmann (1995), Grinblatt and Titman (1992), Gruber (1996), and Hendricks, Patel, and Zeckhauser (1993).

[2] For recent evidence supporting the presence of subgroups of skilled managers, see Avramov and Wermers (2006).

Wermers (2002) finds a substantial level of so-called "style drift" among U.S. mutual funds during the 1975 to 2002 period. As shown by Ferson and Schadt (1996), inferences about manager ability may change substantially when we adjust for the shifting style or risk loadings of a professionally managed portfolio.

Third, holdings data allow an examination of fund manager abilities before expenses and trading costs, which can provide important insights into asset-allocation or security-selection talents. For example, a manager may hold talents in picking securities, but is handicapped by a fund that is too small, making the scale of the fund too expensive to support (through the expense ratio). Identifying such a manager may allow a restructuring of the fund (i.e., through a merger or by creating a new separate account) that overcomes such a handicap while retaining the performance-generating aspects of the fund.

Fourth, an analysis of portfolio holdings adds to our ability to decompose the sources of value-added by a manager. For example, a manager may have strengths in choosing technology stocks, but not telecommunication stocks. Or, the manager's performance may be concentrated in her top 10 holdings, rather than being spread more evenly. Such insights may allow us to better predict whether specific talents in which we are interested are likely to persist.

Finally, benchmarking is more precise when applied on a security-by-security basis, since each security holding constitutes a separate observation about manager ability.[3] Thus, portfolio holdings increase the speed of convergence of the estimated manager ability to her true ability by the law of large numbers.

In the following Section 5, we review several approaches to measuring and attributing performance using portfolio weights. These methods are derived from recently published academic papers, and should be considered as being state-of-the-art; they are also applicable to a wide variety of practical performance evaluation situations. We will also describe extensions and applications of these methods, where appropriate.

2. Unconditional holdings-based performance measurement

A pioneering study that uses portfolio holdings data obtained from periodic SEC filings of mutual funds is Grinblatt and Titman (1989a). This paper breaks new ground by examining performance at the portfolio holdings level, that is, using holdings applied to changes in the closing price quotations of stocks (plus any cash dividends) to compute the return of an asset manager. Grinblatt and Titman label this the "hypothetical performance" of a fund manager, since the manager (or those who would mimic the manager) could not exactly replicate the performance computed from closing prices due to the reality of trading costs.[4] Grinblatt and Titman (1989a) proceed to regress these computed hypothetical returns on those of one or more market-based benchmarks to determine the hypothetical, pre-cost alpha of a fund manager. Therefore, while this paper is the first to directly examine portfolio holdings, the authors do not overcome the benchmark-error problem outlined by Roll (1978).

[3] Although the adding-up constraint implies that the choices of holdings are not truly independent, we can treat them as approximately so since the manager chooses weights for a large set of securities in her portfolio. Note that this constraint only implies a correlation between weights on the order of $1/N$. Thus, even for managers of relatively small portfolios, such as a manager of a biotechnology fund, portfolio weights provide a set of N approximately independent observations of fund manager ability.

[4] Another important complication is that closing prices may not reflect the actual value of securities if sufficient liquidity is not present.

In response to the Roll criticism, academic researchers have built on Grinblatt and Titman (1989a) by developing measures of portfolio performance that allow weights to play a more central role in the formation of the benchmark(s) against which performance is measured. These new, portfolio holdings measures (PHM) are grounded in the concept of performance being measured as the covariance between lagged weights and current returns,

$$\mathrm{PHM}_t = \mathrm{cov}(\tilde{w}_{t-1}, \tilde{R}_t). \tag{1}$$

This concept is very simple: a skilled manager will exhibit portfolio weights that move in the same direction as future returns. Of course, since covariances are unscaled (as opposed, for example, to correlation coefficients), such covariance-based measures assign a higher performance to a manager who is more aggressive; a good manager who aggressively turns over his portfolio will achieve a higher PHM than a great manager who is more cautious.[5] This issue aside, the PHM may be measured in a few alternative ways, simply by the various approaches to defining covariance, that is,

$$\begin{aligned}\mathrm{PHM}_t &= \mathrm{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E\{\tilde{w}_{t-1}[\tilde{R}_t - E(\tilde{R}_t)]\} = E\{[\tilde{w}_{t-1} - E(\tilde{w}_{t-1})]\tilde{R}_t\} \\ &= E\{[\tilde{w}_{t-1} - E(\tilde{w}_{t-1})][\tilde{R}_t - E(\tilde{R}_t)]\}. \end{aligned} \tag{2}$$

These alternative approaches end up being much more than mathematical technicalities in practice. Specifically, these different expansions of covariance imply very different choices of benchmarks for a managed portfolio, and, therefore, have very different implications when applied to practical performance evaluation problems. For example, it may be easier to estimate expected portfolio weights, $E(\tilde{w}_{t-1})$, given available data, than expected security returns, $E(\tilde{R}_t)$ or, vice-versa. Or, if precise estimates of both expected weights and expected returns are readily available, then we would expect that the sample analogue of the expansion, $\mathrm{PHM}_t = E\{[\tilde{w}_{t-1} - E(\tilde{w}_{t-1})][\tilde{R}_t - E(\tilde{R}_t)]\}$ might converge more rapidly to the true performance of a manager than those of the alternative expansions, $\mathrm{PHM}_t = E\{[\tilde{w}_{t-1} - E(\tilde{w}_{t-1})]\tilde{R}_t\}$ or $\mathrm{PHM}_t = E\{\tilde{w}_{t-1}[\tilde{R}_t - E(\tilde{R}_t)]\}$.

Of course, a natural issue that first arises in applying these different approaches is the estimation of $E(\tilde{w}_{t-1})$ and/or $E(\tilde{R}_t)$. The first, the expected (or bogey or normal) weight of a security in a manager's portfolio may be estimated based on its weight in a commonly used benchmark portfolio, such as a market index. This approach might make sense, and may circumvent the Roll benchmark-choice problem, if the manager's explicit mandate is to track such an index, while adding a higher average return than the index with a controlled tracking error around it.[6] Alternatively, the expected weight may be based on the average or cumulative weights of other managers with the same mandate or self-declared investment objective, as proposed by Kandel, Kandel, and Wermers (2005). Or, the expected weight might be entirely based on information in the manager's portfolio holdings—using past or future portfolio weights as expected current portfolio weights.

The second estimation, of $E(\tilde{R}_t)$, the expected return of a security in a manager's portfolio may be estimated based on the average of the security's entire history of returns, its forecasted return based on a factor model, or its single- or multiple-period return during a particular past or future time period (relative to period t, the performance evaluation period). We next illustrate two

[5] We will discuss extensions to pure covariance-based measures that address this issue in a later section of this paper.

[6] However, Roll (1992) shows that managers can game such an explicit benchmark as well, simply by choosing a portfolio with a beta greater than unity with respect to the benchmark.

performance evaluation approaches that, while similar in their mathematical formulation, choose to estimate different quantities in Equation (2)—one chooses to estimate each security's $E(\tilde{R}_t)$, while the other chooses to estimate each security's $E(\tilde{w}_{t-1})$ in a given portfolio.

2.1. The self-benchmarking method of performance evaluation

2.1.1. Statistical foundations

Copeland and Mayers (CM; 1982) and Grinblatt and Titman (GT; 1993) propose new measures of performance that completely disregard standard market benchmarks (such as the Standard and Poor's 500 index) or peer-group benchmarks (such as the average return of same-objective funds) in favor of a "homemade" or bootstrapped benchmark. The CM approach uses the first expansion of the covariance in Eq. (2), $\mathrm{PHM}_t = \mathrm{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E\{\tilde{w}_{t-1}[\tilde{R}_t - E(\tilde{R}_t)]\}$, while the GT approach builds on the second, $\mathrm{PHM}_t = \mathrm{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E\{[\tilde{w}_{t-1} - E(\tilde{R}_t)]\tilde{R}_t\}$. Either of these definitions of covariance should be interpreted as a time-series covariance for a single security holding; CM and GT then propose that a manager's performance might reasonably be measured as the summed covariances across all security holdings, or

$$\mathrm{CM} = \mathrm{GT} = \sum_{j=1}^{N} \mathrm{cov}(\tilde{w}_{j,t-1}, \tilde{R}_{j,t}), \tag{3}$$

which measures the aggregate correctness of the manager's portfolio bets, across all N securities. Of course, the true covariance must be estimated with time-series data, and here is where CM and GT differ. Specifically, CM recommend the estimation,

$$\mathrm{CM} = \sum_{j=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \tilde{w}_{j,t-1}[\tilde{R}_{j,t} - E(\tilde{R}_{j,t})], \tag{4}$$

while GT favor the estimation,

$$\mathrm{GT} = \sum_{j=1}^{N} \frac{1}{T} \sum_{t=1}^{T} [\tilde{w}_{j,t-1} - E(\tilde{w}_{j,t-1})]\tilde{R}_{j,t}. \tag{5}$$

Eq. (5) can be rewritten, for convenience, as

$$\mathrm{GT} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} [\tilde{w}_{j,t-1} - E(\tilde{w}_{j,t-1})]\tilde{R}_{j,t}. \tag{6}$$

This version of the sample measure allows the easy computation of the GT component for a single time period,

$$\mathrm{GT}_t = \sum_{j=1}^{N} [\tilde{w}_{j,t-1} - E(\tilde{w}_{j,t-1})]\tilde{R}_{j,t}, \tag{7}$$

such that $\text{GT} = \frac{1}{T}\sum_{t=1}^{T}\text{GT}_t$. Similarly, the CM component for a single time period is computed as:

$$\text{CM}_t = \sum_{j=1}^{N}\tilde{w}_{j,t-1}[\tilde{R}_{j,t} - E(\tilde{R}_{j,t})], \tag{8}$$

with $\text{CM} = \frac{1}{T}\sum_{t=1}^{T}\text{CM}_t$. It is easily shown that, as T becomes very large, the two measures converge.

The next step is to decide on a proxy for the expected weight for security j at the end of period $t-1$, $E(\tilde{w}_{j,t-1})$, or for the expected return of security j, $E(\tilde{R}_{j,t})$. There are many approaches to this problem, as discussed in the prior section. Grinblatt and Titman (1993) propose that the past weight on security j is the best proxy for the security's expected weight. Specifically, using a market- or peer-based benchmark portfolio may allow the manager to game the benchmark by overweighting (relative to the benchmark) securities with higher expected returns, and underweighting securities with lower expected returns.[7] Further, using the manager's *future* portfolio weight as a proxy for expected current weights would bias the GT measure if the manager implements trading strategies that condition on past security returns. For instance, a manager who overweights securities with high past returns (a relative strength or momentum trader), will exhibit a future weight that is correlated with current returns—thus, spuriously reducing the estimated performance of the manager. For these reasons, GT recommend the past weight of a security in the manager's portfolio as a proxy for current expected, or bogey weight.

For the same reason, CM choose the future return of a security as a proxy for its expected return, $E(\tilde{R}_{j,t})$. With these proxy choices, note that the GT expansion is advantageous, relative to the CM expansion, since it does not require a security to survive past the current period, t, to be included in the current period's performance calculation (recall that the CM expansion requires the future return of a security for the current period computation). Nevertheless, the GT measure has its costs – it requires that the manager must exist for one full period before we can measure performance, while the CM requires only a single observation of the manager's weights – at the end of date $t-1$. In most practical situations, we would expect that managers have existed for at least 1 year before we wish to evaluate them, thus, the GT measure is preferred in most situations. As such, we proceed with the GT measure, although the application of the CM measure follows similar logic.[8]

At this point, we note another nice feature of the GT measure: it is exactly equal to the difference between the next-period return earned by the current portfolio and a historical portfolio held by a manager, or

$$\text{GT}_t = \sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})\tilde{R}_{j,t}, \tag{9}$$

where $\tilde{w}_{j,t-1}$ is the end-of-month $t-1$ (beginning of month t) portfolio weight of stock j held by the manager, $\tilde{w}_{j,t-k-1}$ is the weight of the same stock lagged k months, and $\tilde{R}_{j,t}$ is

[7] Although a manager could also move toward higher expected return securities over time, thus gaming the past-weight benchmark, such a strategy would be much more difficult to sustain over time unless the manager overweights temporarily high-risk securities.

[8] In some cases, if historical portfolio weights and future returns are readily available for all securities held by a manager, the third expansion of Eq. (2) might be used. If so, the application of this approach would, again, be similar to that shown for the GT and CM measures.

the month-t return of stock j. Note that, with this measure, the benchmark used to adjust the return of a portfolio for its risk in a given month is the current month's return earned by the portfolio (held by the same manager) k months prior to the current month's holdings. Taking this interpretation further, the time-series average GT measure, $\text{GT} = \frac{1}{T-k}\sum_{t=k+1}^{T}\text{GT}_t$, represents the mean return of a zero-investment portfolio—long the current portfolio and short the historical portfolio of a given manager. Consider the risk implications of the GT measure: if the systematic risks of the current and benchmark portfolios are the same (plus random noise) for a given manager having no selectivity or timing abilities [as defined by Grinblatt and Titman (1989b)], then the average portfolio represented by GT will have diminishing systematic risk, and this average portfolio return will converge to zero as T increases—in small samples, both risk and return will (under certain reasonable conditions) be insignificant for that manager.

In addition, while prior performance measures are susceptible to benchmark error (as in Roll (1978)) and other forms of model misspecification, these errors are much less problematic with the GT measure. This results from the form of the GT measure—it differences portfolio returns, so any misspecification that remains is due only to differences in loadings (on any omitted risk factors) between the current and historical portfolios. In effect, past holdings represent the normal or bogey risks taken by a particular manager.

Therefore, this measure will be biased only if a manager tilts toward stocks having temporarily high risk loadings. In such a case, the current portfolio will have a higher systematic risk than the historical portfolio, and the manager will exhibit a GT measure that is an upward-biased estimate of manager talent. An approach that can help to correct for such biases in the GT measure is to regress the time-series, GT_t, on standard benchmarks. For example, for U.S. stock portfolios; the following four-factor regression can be run:

$$\text{GT}_t = \alpha + b \cdot \text{RMRF}_t + s \cdot \text{SMB}_t + h \cdot \text{HML}_t + u \cdot \text{UMD}_t + e_t\,, \tag{10}$$

where RMRF_t equals the month-t excess return on a value-weighted aggregate market proxy portfolio; and SMB_t, HML_t, and UMD_t equal the month-t returns on value-weighted, zero-investment factor-mimicking portfolios for size, book-to-market equity, and 1-year momentum in stock returns.[9]

The remaining benchmark-adjusted GT measure, or α, is the manager stockpicking talent, adjusted for any strategy employed by a manager (or, any passive movement in portfolio weights over time) that tilts toward stocks with temporarily high loadings on RMRF, SMB, HML, or UMD. An example is a manager who, by chance or by choice, adds stocks with high 1-year past returns to her portfolio. The current portfolio will have a higher loading on UMD than the historical portfolio, so GT_t will exhibit a positive u coefficient in regression (10)—but α in this regression will be cleaned of the influence of such a risk loading.

Note, as mentioned, that the GT approach has a cost: in this case, the portfolio holdings for the first k months must be set aside as benchmarks for future portfolios. Therefore, the performance of the manager is not available during these first k months (which also eliminates short-lived funds), making this choice of k important.[10] The choice of k has other implications for this measure as

[9] These research returns are available (and updated) via Professor Kenneth French's website at Dartmouth University (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/).

[10] However, in some cases, various alternative benchmarks might be applied to recover these first k months, such as using peer-group weights.

Table 1
GT performance measures for 155 U.S. domestic equity funds, 1975–1984

	No. of funds	Performance measure					
		Lagged 3 months			Lagged 12 months		
		Mean performance	t-statistic	Wilcoxian probability	Mean performance	t-statistic	Wilcoxian probability
Total sample	155	.37	1.47	.233	2.04	3.16*	.004
Aggressive growth	45	.39	.98	.475	3.40	3.55*	.004
Balanced	10	−.48	−1.87	.057	.01	.03	.902
Growth	44	.66	2.01*	.017	2.41	2.94*	.009
Growth-income	37	.14	.61	.095	.83	1.75	.107
Income	13	.54	1.54	.475	1.33	2.64*	.002
Special purpose	3	−.10	−.16	.233	.21	.19	.711
Venture capital/special situation	3	1.26	1.07	.812	2.66	1.43	.035

* Type I error < .05.

well. If k is chosen to be too small, we will eliminate performance that occurs beyond the first k months that a manager holds a stock, since, at that point, the difference in weights from Eq. (9) will be zero if the manager holds the stock in a constant amount during these k months; if k is too large, it is more likely that the measure may include some systematic risk differences in the manager's portfolio over these longer time periods.

When applying this measure to U.S. domestic equity mutual funds, Grinblatt and Titman (1993) find that fund performance appears to increase when varying k from 1 through 12 months, beyond which further increases in fund performance are small. Thus, they recommend a 12-month portfolio lag for these U.S. funds. A similar approach may be used to choose a reasonable value of k in other applications—increasing the chosen k (within reasonable ranges) until further performance changes appear to be negligible or add more noise (increase in standard deviation) than signal (increase in point estimate).

2.1.2. Empirical evidence

Table 1 shows the results of the GT measure applied to 155 U.S. domestic equity mutual funds that exist for the entire period from 1975 to 1984.[11] Time-series averages (GT) of two variants of the GT measure of Eq. (9) are presented: the first uses a lag of $k = 3$ months, while the second uses a lag of $k = 12$ months. Note that the measure using the 12-month lag exhibits over four times the level of performance for the average fund (2.04%) as the 3-month lag (0.37%), indicating that the portfolio choices of the average manager exhibit substantial levels of performance during the remaining nine months. Thus, using a 3-month lag would result in an underestimation of the skills of the manager. Note, also, that a simple time-series t-statistic is used to determine the statistical precision of the point estimate, since the GT measures for each time period, given by Eq. (9) are independent under the null hypothesis of no manager ability. For comparison, Wilcoxon non-parametric rank statistic p-values are also shown.

[11] This table is reproduced from Grinblatt and Titman (1993).

Wermers (1997) updates the GT study to include all funds existing between 1975 and 1994, inclusive. Wermers finds that the average GT measure of U.S. domestic equity funds (equally weighted over all survivors and non-survivors, rebalanced at the beginning of each year) over this period is 1.7% per year, using the 1-year lagged portfolio weights ($k=12$), which is similar to the Grinblatt and Titman (1993) result (2.04% per year) shown in Table 1.

2.1.3. Relation to the Brinson, Hood, and Beebower attribution approach

This GT measure also holds some statistical advantages over standard applications of the Brinson, Hood, and Beebower (BHB; 1986) attribution formulation. Commonly used applications of the BHB approach generally involve choosing a market-based index as a benchmark for a manager. Then, taking a pure equity portfolio manager as an example, the stock selection ability of this manager is computed as:

$$\mathrm{BHB}_t = \sum_{j=1}^{N} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - \tilde{R}_t^{\text{index}}), \tag{11}$$

where $\tilde{R}_t^{\text{index}}$ is the return on the benchmark index during month t. Related to our discussion above, if the index is misspecified for this manager (or the proxy for the index is poorly chosen), then the BHB measure will contain a bias that is directly proportional to the loading of the manager on any omitted factors. By contrast, as discussed above, the GT measure will contain a bias that is proportional to the difference in loadings between the current and historical portfolios, which will be much smaller in most applications.

In addition, the BHB measure uses a market index chosen to match the manager's stated investment objective at the beginning of the evaluation period—which, clearly, depends on the truthful revelation of such an objective. Thus, the manager might choose to game this measure by choosing an index that is expected to underperform his actual chosen portfolio. On the other hand, the GT measure uses an actual past portfolio of a manager as a proxy for the manager's current strategy—making it more difficult for a manager to game the measure for any substantial period of time.[12]

Fortunately, the GT approach can be integrated into a BHB-type attribution system by substituting a historical portfolio return for a market-based index:

$$\mathrm{BHB}_t^{\mathrm{GT}} = \sum_{j=1}^{N} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - \tilde{R}_t^{\text{historical}}), \tag{12}$$

where $\tilde{R}_t^{\text{historical}} = \sum_{j=1}^{N} \tilde{w}_{j,\text{historical}} \tilde{R}_{j,t}$.

The GT measure can also be decomposed into style bets, style timing bets, market timing bets, industry bets, etc., as well as the remaining residual stockpicking bets by simple decompositions of the sum in Eq. (9). For example, if we wish to decompose into the overall performance

[12] As mentioned above, however, the manager can game the GT measure if stocks have substantial time-variation in expected returns that are predictable, such as the 1-year momentum effect (see Grinblatt, Titman, and Wermers (1995)). But, a chosen style index used with the BHB method would likely be even more susceptible to such gaming.

from industry bets and the remaining residual stock selectivity bets, we would decompose as,

$$\mathrm{GT}_t = \underbrace{\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})\tilde{R}_t^{\mathrm{IND}(j)}}_{\text{Industry Performance Component}} + \underbrace{\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})(\tilde{R}_{j,t} - \tilde{R}_t^{\mathrm{IND}(j)})}_{\text{Stock Selectivity Component}}, \qquad (13)$$

where $\tilde{R}_t^{\mathrm{IND}(j)}$ is the month-t return of the industry portfolio to which stock j belongs at the beginning of the month. As with BHB, Eq. (13) can be decomposed into further partial sums, as desired, to explore the detailed sources of manager returns. Each partial sum is then averaged over all months. Or, if desired, each partial sum can be averaged only over certain months (such as January) or certain time-ranges (such as 3-year periods) to analyze the time-series variation in manager performance.

In this section, we have argued that the GT measure holds significant advantages over standard approaches that use market indexes as benchmarks since the GT approach uses lagged manager weights to form dynamic benchmarks. However, we may wish to use our knowledge of the factors or characteristics that, from past research, are known to drive security returns in order to obtain a more precise measure of performance. For instance, as mentioned above, a manager can potentially game the GT measure by overweighting securities with temporarily high levels of risk. To overcome such concerns, we must define the sources of risk in our security universe—in effect, the cost of increased performance evaluation precision is that we must specify the main influences on security returns for all securities in the investable universe of a manager.[13] For some asset managers, this may not be practical—thus, the GT approach may be preferred. However, in the case of U.S. domestic equities, extensive research has documented the most important influences on returns. This research has identified the drivers of cross-sectional differences in stock returns as well as time-series variation in the returns of a given stock. We next describe an approach for evaluating equity portfolios in the U.S. – with a discussion on extensions in non-U.S. equity markets – that uses the results of this past research.

2.2. *The DGTW method of performance evaluation for equity portfolios*

The Daniel, Grinblatt, Titman, and Wermers (DGTW; 1997) approach applies the results of prior empirical research on the factors that drive stock returns. This research, which includes Fama and French (1992, 1993, 1996) and Jegadeesh and Titman (1993), shows that the market index as well as indexes that proxy for the size, book-to-market, and momentum effects are sufficient to explain the vast majority of the cross-sectional and time-series variation in U.S. stock returns. International evidence, such as Rouwenhorst (1998), indicates that similar factors also explain non-U.S. stock returns.

[13] This is fairly straightforward for U.S. domestic equities, and (with recent research) for non-U.S. equities in many developed countries. However, the factors that drive bond returns as well as hedge fund returns are not clearly understood; thus, it is difficult to unambiguously create benchmark portfolios for these markets. In addition, a manager holding a mixed portfolio of both stocks and bonds presents a problem when using defined benchmarks developed for the stock market, although some approaches exist.

The DGTW portfolio holdings-based measures include three subcomponents:

1. the current portfolio-weighted return on stocks currently held by the fund, in excess of returns (during the same time period) on matched control portfolios having the same style characteristics (selectivity- or stockpicking-based returns);
2. the current portfolio-weighted return on control portfolios having the same characteristics as stocks currently held by the fund, in excess of time-series average returns on those control portfolios (style-timing-based returns); and
3. the time-series average returns on control portfolios having the same characteristics as stocks currently held (style-based returns).

These three components decompose the return on the portfolio holdings of a fund; thus, they attribute returns before any trading costs or expenses are considered. We describe these three components next.[14]

2.2.1. The characteristic selectivity (CS) measure

The first component of performance measures the stock-picking ability of the fund manager, controlling for the particular style used by that manager. This measure of stock-picking ability, which is called the characteristic selectivity measure (CS), is computed during month t as,

$$\mathrm{CS}_t = \sum_{j=1}^{N} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - \tilde{R}_t^{b_{j,t-1}}), \tag{14}$$

where $\tilde{w}_{j,t-1}$ is the portfolio weight on stock j at the end of month $t-1$, $\tilde{R}_{j,t}$ is the month-t buy-and-hold return of stock j, and $\tilde{R}_t^{b_{j,t-1}}$ is the month-t buy-and-hold return of a value-weighted portfolio that is matched to stock j based on its characteristics at the end of month $t-1$.

To construct the characteristic-matched benchmark portfolio for a given stock at the beginning of a given month, we characterize that stock over three dimensions—the market capitalization of equity (size), the ratio of book-value of equity to market-value of equity, and the prior-year return. Forming these matching portfolios proceeds as follows—this procedure is based on DGTW, and is described in more detail in that paper. First, all stocks (listed on NYSE, AMEX, or Nasdaq) having book-value of equity information in Compustat, and stock return and market capitalization of equity data in the Center for Research in Security Prices (CRSP) stock files, are ranked, at the end of each June, by their market capitalization. Quintile portfolios are formed (using NYSE size quintile breakpoints), and each quintile portfolio is further subdivided into book-to-market quintiles, based on their book-to-market data as of the end of the December immediately prior to the ranking year. Finally, each of the resulting 25 fractile portfolios are further subdivided into quintiles based on the 12-month past return of stocks through the end of May of the ranking year. This three-way ranking procedure results in 125 fractile portfolios, each having a distinct combination of size, book-to-market, and momentum characteristics.[15] The three-way ranking

[14] These measures are developed in DGTW, and are more fully described there. In that paper, the authors argue that decomposing performance with the use of benchmark portfolios matched to stocks on the basis of the size, book-to-market, and prior-year return characteristics of the stocks is a more precise method of controlling for style-based returns than the method of decomposing performance using net returns in factor-based regressions, such as those used by Carhart (1997).

[15] Thus, a stock belonging to size portfolio one, book-to-market portfolio one, and prior return portfolio one is a small, low book-to-market (growth) stock having a low prior-year return.

procedure is repeated at the end of June of each year, and the 125 portfolios are reconstituted at that date.

Value-weighted returns are computed for each of the 125 fractile portfolios, and the benchmark for each stock during a given month is the buy-and-hold return of the fractile portfolio of which that stock is a member during that month. Therefore, the characteristic-adjusted return for a given stock is computed as the buy-and-hold stock return minus the buy-and-hold (value-weighted) matched benchmark return during the same month.[16] Finally, the characteristic selectivity measure of the stock portfolio of a given mutual fund during month t, CS_t, is computed as the portfolio-weighted characteristic-adjusted return of the component stocks in the portfolio, where the stock portfolio is normalized so that the weights add to one (to focus on the performance of the equity portion of the portfolio).

A caveat is in order regarding the interpretation of the CS measure, as it controls for only three characteristic dimensions of stock—size, book-to-market, and past returns. Recent research has shown that mutual funds show a distinct preference for other stock characteristics that are related to average returns—for example, stocks with greater liquidity (see Chen, Jegadeesh, and Wermers (2000)).[17] For example, one might argue that our CS measure underestimates the stock-picking talents of funds, since we do not control for the lower average returns that accrue to stocks with greater liquidity. These potential missing factors probably do not impact the practical uses of the DGTW approach, but the literature has yet to fully explore such issues.

2.2.2. The characteristic timing (CT) measure

The above stock-selectivity measure does not capture the ability of the fund manager to time the various stock characteristics. Indeed, fund managers can generate additional performance if size, book-to-market, or momentum strategies have time-varying expected returns that the manager can exploit by "tilting" portfolio weights toward stocks having these characteristics when the returns on the characteristics are highest. Thus, our second component of performance measures a fund manager's success at timing the different stock characteristics; this component is termed the "characteristic timing" (CT) measure. The month t component of this measure is,

$$\mathrm{CT}_t = \sum_{j=1}^{N} (\tilde{w}_{j,t-1} \tilde{R}_t^{b_{j,t-1}} - \tilde{w}_{j,t-k-1} \tilde{R}_t^{b_{j,t-k-1}}). \tag{15}$$

Note that this expression deducts the month-t return of the month-$(t-k-1)$ matching characteristic portfolio for stock j (times the portfolio weight at $t-k-1$) from the month-t return of the month-$(t-1)$ matching characteristic portfolio for stock j (times the portfolio weight at $t-1$). Thus, a fund manager who increases the fund's weight on stock j just before the payoff to the characteristics of stock j is highest will exhibit a large CT measure. The choice of k, of course, will depend on the frequency used by the manager to time characteristic-based returns among stocks. For instance, if such timing strategies generally pay off within 1 year, then we might choose $k=12$, thus, lagging weights 12 months from the current portfolio date, $t-1$. This choice of k is used by DGTW to evaluate U.S. domestic equity mutual funds.

[16] The composition of these benchmark portfolios, as well as their returns are available via http://www.smith.umd.edu/faculty/rwermers/ftpsite/Dgtw/coverpage.

[17] See Lee and Swaminathan (2000) and Datar, Naik, and Radcliffe (1998) for evidence that more liquid stocks earn lower average returns.

2.2.3. The average style (AS) measure

To measure the returns earned by a fund because of that fund's tendency to hold stocks with certain characteristics, we employ our third performance component, the average style (AS) return measure. The month-t component of this measure is

$$\mathrm{AS}_t = \sum_{j=1}^{N} \tilde{w}_{j,t-k-1} \tilde{R}_t^{b_{j,t-k-1}}. \tag{16}$$

Each stock held by a fund at the end of month $t-k$ is matched with its characteristic-based benchmark portfolio of that date. The month-t return of this benchmark portfolio is then multiplied by the end of month $t-k-1$ portfolio weight of the stock, and the resulting product is summed over all stocks held by the fund at the end of month $t-k-1$ to give the month-t AS component. Note that, by lagging weights and benchmark portfolios by k-periods, we eliminate returns due to the timing effect described by Eq. (15). For example, a fund that successfully buys high book-to-market stocks when returns to such a strategy are unusually high will not exhibit an unusually high AS return, since this strategy will most likely involve moving into stocks shortly before the unusually high book-to-market return. However, a fund that systematically holds high book-to-market stocks to boost its portfolio return (without trying to time the effect) will exhibit a high AS return. Note that k should be chosen to be the same value in both Eqs. (15) and (16).

The AS measure of a fund may differ from the return on a broad market index for a couple reasons. First, the AS measure may contain a compensation for the fund loading on covariance-based risk factors that differs from the market portfolio's loadings. And, second, the AS measure may contain return premia for the fund loading on non-covariance-based characteristic factors. We do not attempt to separate these two sources of AS return premia here, but we note the phenomena and leave the interpretation to the reader.

2.2.4. Summing the components

Note that the sum of the CS, CT, and AS measures equals the total portfolio-weighted (pre-expense and pre-trade cost) return on the stockholdings of a given fund (we also call this the gross return of the fund). That is,

$$\mathrm{GR}_t = \mathrm{CS}_t + \mathrm{CT}_t + \mathrm{AS}_t. \tag{17}$$

Note, also, that computations of the AS and CT measures begin after a lag of k-periods, as we must use k-period lagged portfolio weights to compute these measures.

2.2.5. Comparison of DGTW measures with factor-based regression approaches

One might reasonably ask why we cannot, more easily, use sophisticated multi-factor returns-based performance evaluation methods to arrive at similar measures—these, of course, do not require gathering and processing portfolio weights! For example, Carhart (1997) develops a four-factor regression method for estimating mutual fund performance. This four-factor model is based on an extension of the Fama and French (1993) factor model, and is described as,

$$R_{j,t} - R_{F,t} = \alpha_j + b_j \cdot \mathrm{RMRF}_t + s_j \cdot \mathrm{SMB}_t + h_j \cdot \mathrm{HML}_t + u_j \cdot \mathrm{UMD}_t + e_{j,t}. \tag{18}$$

Here, $R_{j,t} - R_{F,t}$ equals the excess net return of fund j during month t (the fund net return minus T-bills); RMRF_t equals the month-t excess return on a value-weighted aggregate market proxy portfolio; and SMB_t, HML_t, and UMD_t equal the month-t returns on value-weighted, zero-investment factor-mimicking portfolios for size, book-to-market equity, and 1-year momentum in stock returns.

While the Carhart regression approach seemingly controls for known sources of systematic time-series and cross-sectional return variability among U.S. equities, Daniel and Titman (1997) show strong empirical evidence that factor-based risk controls can be beaten with mechanical strategies that, for example, pick certain value stocks. The point of Daniel and Titman is that such strategies will exhibit a positive alpha in a Carhart regression, as long as we choose a portfolio of stocks (e.g., value) that have a low covariance with the factors (e.g., HML in the Carhart model) that are designed to capture their common return influences (value). Daniel and Titman find that stocks with value characteristics, such as those having a high book-value to market-value of equity ratio, generate superior returns to stocks that have growth characteristics, even if the value stocks have a low covariation with other value stocks. This finding implies that investors do not bid down the prices of value stocks because they have an exposure to an unknown risk factor (such as financial distress), but that they irrationally do so, perhaps due to over-extrapolating recent poor earnings results for such stocks.

Regardless of the interpretation of the Daniel and Titman results, the DGTW approach uses the Daniel and Titman results to benchmark stocks against their characteristic peers (e.g., other high book-to-market stocks) rather than against other stocks with which they have a high return covariation. Practically speaking, this means that the DGTW-measured performance controls more accurately for simple, mechanical strategies for which we may not wish to reward a fund manager (since many index funds provide bets on these strategies for low cost). Nevertheless, the limited data requirements of the regression approach make it appealing in many practical cases. Therefore, it is important to compare the DGTW and Fama-French/Carhart approaches to performance evaluation and attribution.

Table 2 shows the relation between the DGTW components and the Fama-French/Carhart components. One should note that the dynamic style control approach of DGTW allows for several advantages over the fixed style coefficient regression approach of Fama-French/Carhart. The most important of these advantages are the smaller standard error for the selectivity component, CS, relative to α, and the (correspondingly) more precise measure of returns attributable to style loadings, AS, relative to the sum of the fixed coefficients times average style return premia of F-F/Carhart. In addition, the Carhart α can be biased if style loadings by a fund manager are changing over time.

2.2.6. Extensions

As with the GT performance measure of Eq. (9), the CS performance measure of Eq. (14) can easily be decomposed into partial sums. This decomposition is a useful technique to determine the

Table 2
Comparison of attribution by DGTW and Fama-French/Carhart approaches

	Selectivity	Style timing	Average style
DGTW component	CS	CT	AS
FF/Carhart component	α	–	$b \cdot \overline{\text{RMRF}} + s \cdot \overline{\text{SMB}} + h \cdot \overline{\text{HML}} + u \cdot \overline{\text{UMD}}$
Comments	CS has lower standard errors than α	Assumed constant factor loadings of FF/Carhart does not allow timing measures	Dynamic style measurement of AS allows more precise measure than assumed constant b, s, h, and u of FF/Carhart

attribution of performance in a manager's portfolio. For instance, Kacperczyk, Sialm, and Zheng (2005) find that mutual funds with concentrated holdings in a few industries outperform those with more diverse holdings—suggesting that manager skills are industry-specific. If so, we may wish to measure the performance of stocks held by the manager within a certain industry, ignoring the other holdings of that manager. Alternatively, we may wish to measure the performance of the manager's top 10 holdings, regardless of their sector membership. Further, Wermers (2005) shows that stocks purchased by U.S. mutual fund managers in response to strong cash inflows from investors outperform the other holdings of the managers. In all of these cases, a modification of the performance measures may be used to capture the performance of the subportfolio of interest.

2.2.6.1. Attributing subportfolio performance. To illustrate, the characteristic selectivity measure is modified as follows. Suppose that we are interested in measuring the performance of the subportfolio of stocks belonging to set S, where S refers, for example, to a certain sector of the market (e.g., technology stocks). The CS measure is decomposed to measure the partial sum corresponding to stocks within set S. That is, the selectivity performance of the manager that is attributable to holding stocks within set S is measured as:

$$\mathrm{CS}_{S,t} = \sum_{j \in S} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - \tilde{R}_t^{b_{j,t-1}}). \tag{19}$$

A different question may also be addressed: how skilled (in theory) would this manager be if she were constrained to buy only stocks within set S? A further modification that normalizes set S portfolio weights to add to unity allows an answer:

$$\mathrm{CS}'_{S,t} = \frac{\sum_{j \in S} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - \tilde{R}_t^{b_{j,t-1}})}{\sum_{j \in S} \tilde{w}_{j,t-1}}. \tag{20}$$

2.2.6.2. Normalizing for turnover. An ideal property of a performance measure is to rank managers by the precision of their private information, or simply put, skills that cannot be captured through simple strategies that may be implemented by an uninformed observer. As discussed after Eq. (1), covariance-based measures assign a higher performance measure to managers implementing higher levels of portfolio turnover, holding constant their actual skills in forecasting security returns. Such differences in turnover can arise from the agency problems inherent in asset management, since most managers own a very small fraction of their managed portfolios. Such agency problems include compensation that is insufficiently sensitive to performance, or the labor-market concerns of a manager (i.e., the manager's reputational concerns or concerns over being fired). These agency problems, in theory, should not affect an investor's performance evaluation, since an investor can scale investments (financed through borrowing or short-selling other investments) to any level to increase or decrease the return desired. However, covariance-based measures, by their very nature, are susceptible to turnover-related biases.[18]

In general, it is very difficult to properly adjust for the risk-aversion of a manager, as the manager may not view all investment opportunities with the same level of risk-aversion. For example, a

[18] It is important to note that returns-based performance measures are also plagued with the risk-aversion problem. One approach to adjusting such measures is to normalize (divide) returns-based regression alphas by the turnover of the manager.

durable-goods manager may be more willing to take a large position in an automobile stock relative to a technology stock with similar prospects. Thus, adjusting the covariance measure would require the knowledge of the manager's risk-aversion with respect to all potential investments.

However, we can easily normalize portfolio-based performance measures for risk-aversion with a simplifying assumption, that a manager views all stocks similarly. This assumption can be relaxed when measuring the performance of subportfolios, as described in the previous section—as such subportfolios likely have much more homogeneous risk characteristics, from the viewpoint of the manager, than securities within the overall portfolio.

To adjust for turnover, the investor can implement a turnover-adjusted version of the portfolio holdings-based measure. To illustrate, the turnover-adjusted GT measure during period t is computed as:

$$\mathrm{TAGT}_t = \frac{\mathrm{GT}_t}{\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})}. \tag{21}$$

Similarly, the CS measure, adjusted for turnover, is computed as

$$\mathrm{TACS}_t = \frac{\mathrm{CS}_t}{\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})}. \tag{22}$$

2.2.7. Empirical evidence

The CS measure has been widely applied to evaluate the performance of managed portfolios of securities.[19] As mentioned previously, Kacperczyk et al. (2005) find that mutual funds with concentrated holdings in a few industries outperform those with more diverse holdings—suggesting that manager skills are industry-specific. These researchers find that the decile of U.S. mutual funds that hold the most industry-concentrated portfolios outperform the decile holding the most industry-diversified portfolios by 1.6% per year, according to the difference in CS measures.

In recent years, researchers have also applied the CS methodology to equity portfolios outside the U.S. by forming the characteristic-based benchmarks in the appropriate country stock market. For example, Wylie (2005) follows DGTW (1997) closely in forming characteristic-based benchmarks for the UK market. The procedure is as follows. At the end of June of each year, all London Stock Exchange-listed stocks are partitioned into 125 benchmark portfolios by repeated division into quintiles on the basis of market capitalization at the end of June, then book-to-market ratio during the most recent fiscal year-end, then prior 12-month return (ending in May rather than June to avoid the 1-month return reversal described in Jegadeesh (1990)).[20] Then, the market capitalization-weighted average return of each benchmark portfolio is calculated for each of the 12 months following June to compute benchmark returns for each month of that year. Wylie

[19] The DGTW benchmarks, on which the CS measure is built, have been applied to measure the performance of portfolios of stocks formed as part of an event study, as well as the more conventional application of the benchmarks to measure the performance of an asset manager. For instance, Jeng, Metrick, and Zeckhauser (2003) use the benchmarks to measure the abnormal returns gained by insiders when they trade their company's stock, while Moskowitz and Grinblatt (1999) use the benchmarks to investigate the role of industry in explaining momentum in U.S. equity returns.

[20] Book value is defined as ordinary share capital plus reserves plus deferred and future taxation. Stocks are excluded from the portfolio formation unless they have market capitalization data for June of the current year and December of the preceding year; book value data for the previous 2 years, at least six monthly returns in the year preceding May 31, and are ordinary shares.

(2005) finds that stocks that UK fund managers herd into (buy as a group) underperform those they herd out of by 0.8% during the following year, a result that is opposite that for U.S. stocks (Wermers (1999)).

Iihara, Kato, and Tokunaga (2004) form characteristic-based portfolios from stocks listed on the Tokyo Stock Exchange to study momentum in Japanese stocks. At the end of each June from 1975–1997, all TSE stocks are sorted into five equal groups from small to large based upon their market capitalization. They independently sort TSE stocks into five equal book-to-market (BTM) groups, where BTM is equal to the ratio of book-value to market-value of equity at the end of June for each year. Finally, 25 benchmark portfolios are created from the intersection of the five size and five book-to-market groups. Monthly equal-weighted returns for each of these 25 portfolios are calculated from July of year t to June of year $t+1$.

Other papers that form characteristic-based benchmarks, in the spirit of the DGTW approach, include Pinnuck (2003), who forms benchmarks in Australia. In addition, Jern (2002) forms benchmarks for Finnish fund managers, while Nitibhon, Tirapat, and Wermers (2005) form benchmarks for Thai equity fund managers. Given the popularity of the DGTW approach, it is likely that further developments in other markets will make these benchmarks more widely available to both academics and practitioners. In addition, once the risk factors in fixed-income markets become widely accepted, it is likely that a similar approach will be feasible for measuring the performance of fixed-income asset managers.

2.2.8. *The correlation between performance measures*

An interesting issue is the correlation, across funds, of performance at the stockholdings level (DGTW) and performance at the net returns level (F-F/Carhart). Although, at first blush, it would seem that the correlation would be very high, the issue becomes more interesting if the expenses and transactions costs of funds are also positively correlated with their pre-expense performance. The tendency of funds with superior stock-picking skills to incur higher costs would be consistent with the equilibrium model of Grossman and Stiglitz (1980), where the returns to information gathering and processing skills are equal to the costs. Analyzing both the portfolio holdings and the net returns of funds allows deeper insight into an accounting of the sources and uses of returns earned by a portfolio manager.

Table 3 investigates this issue by presenting cross-sectional correlations (across funds) between various measures of performance.[21] These performance measures are computed over the entire life of each U.S. domestic equity mutual fund during the 1975 to 1994 period. The only restriction we place on a fund to be included in these correlations is that the fund must have at least 24 valid monthly return observations (both for stockholdings and net returns) in order to provide a reasonable degrees-of-freedom in the regression-based measures.

Panel A of the table presents Pearson correlations between three measures of performance at the stock portfolio level: the characteristic selectivity measure (CS), the Carhart measure using the time-series of excess (hypothetical) monthly returns on the stock portfolio as the explained variable ($\alpha^{\text{Gross}}_{\text{Carhart}}$), and the Jensen measure using the same explained variable ($\alpha^{\text{Gross}}_{\text{Jensen}}$). In addition, two measures of performance at the net return level are included: the Carhart measure using the time-series of excess monthly net returns as the explained variable ($\alpha^{\text{Net}}_{\text{Carhart}}$) and the Jensen measure using the same explained variable ($\alpha^{\text{Net}}_{\text{Jensen}}$). Panel B presents Spearman rank-correlations between all of these performance measures.

[21] This table is replicated from Wermers (2000).

Table 3
Correlations between performance measures

Panel A: Pearson correlations

ρ_{Pearson}	CS	$\alpha^{\text{Gross}}_{\text{Carhart}}$	$\alpha^{\text{Net}}_{\text{Carhart}}$	$\alpha^{\text{Gross}}_{\text{Jensen}}$	$\alpha^{\text{Net}}_{\text{Jensen}}$
CS	1	–	–	–	–
$\alpha^{\text{Gross}}_{\text{Carhart}}$	0.57	1	–	–	–
$\alpha^{\text{Net}}_{\text{Carhart}}$	0.36	0.62	1	–	–
$\alpha^{\text{Gross}}_{\text{Jensen}}$	0.58	0.74	0.43	1	–
$\alpha^{\text{Net}}_{\text{Jensen}}$	0.33	0.49	0.84	0.63	1

Panel B: Spearman rank-correlations

ρ_{Spearman}	CS	$\alpha^{\text{Gross}}_{\text{Carhart}}$	$\alpha^{\text{Net}}_{\text{Carhart}}$	$\alpha^{\text{Gross}}_{\text{Jensen}}$	$\alpha^{\text{Net}}_{\text{Jensen}}$
CS	1	–	–	–	–
$\alpha^{\text{Gross}}_{\text{Carhart}}$	0.65	1	–	–	–
$\alpha^{\text{Net}}_{\text{Carhart}}$	0.46	0.63	1	–	–
$\alpha^{\text{Gross}}_{\text{Jensen}}$	0.53	0.67	0.44	1	–
$\alpha^{\text{Net}}_{\text{Jensen}}$	0.40	0.49	0.80	0.68	1

Several observations may be drawn from the two correlation matrices. First, the cross-sectional Pearson correlation (at the gross return level) between the Jensen and Carhart measures is 0.74, which indicates that adding the size, book-to-market, and momentum factors provides a modest increase in the precision of the performance estimate. This result indicates that mutual fund return premia based on loadings on these omitted variables in the Jensen regression are correlated with the intercept of the Carhart regression. The Spearman rank-correlation between these two measures is similar, 0.67.

The correlations between the CS measure and the Carhart and Jensen gross performance measures are 0.57 and 0.58, respectively. These lower correlations support the idea that the CS measure provides more precise adjustments for characteristic-based returns than the regression-based methods.[22] Again, Spearman rank-correlations are similar.

At the net return level, the Carhart and Jensen performance measures are again highly correlated, both with the Pearson correlation and with the non-parametric Spearman rank-correlation. A comparison of these measures at the gross stockholdings return level and the net return level provides further insight. For example, the Carhart measure of the stockholdings of funds is highly correlated with the Carhart net return performance measure. The Pearson correlation is 0.62, while the Spearman correlation is 0.63. These high correlations between gross and net performance indicate that the level of mutual fund expenses and transactions costs, while possibly correlated with fund performance at the stockholdings level, do not eliminate the higher benchmark-adjusted net returns provided by funds with stock-picking talents.

[22] See DGTW for evidence that further supports the use of the CS measure versus the Carhart and Jensen alphas. The CS performance estimates are roughly the same magnitude as the Carhart and Jensen alphas, but the estimated CS standard errors are much lower.

3. Conditional holdings-based performance measurement

The above unconditional performance measures, while straightforward to compute, may be subject to some non-trivial biases. These biases will occur if the manager of the measured portfolio uses macroeconomic information to (dynamically) sharpen her estimates of expected security returns over time. In such a case, the manager can game the unconditional performance measure by overweighting securities that, under current macroeconomic conditions, have higher-than-normal expected returns, and underweighting securities that have lower-than-normal expected returns. Such a strategy will exhibit a positive unconditional performance measure, since the manager's weights will be correlated with returns over time even in the absence of any private information (i.e., information not based purely on widely available macroeconomic information) about expected stock returns. Prior research by Keim and Stambaugh (1986) and Fama and French (1989) indicates that instruments that are useful in predicting future U.S. marketwide equity returns include the: (i) aggregate dividend yield; (ii) default spread; (iii) term spread; and (iv) yield on the three-month T-bill.[23] Further, Ferson and Harvey (1999) find that these variables are relevant in explaining the cross-section of U.S. stock returns.

For instance, suppose that a particular lagged macroeconomic factor (e.g., the T-bill yield) is positively correlated with the next-period return in one sector of the stock market (durable-goods stocks), while inversely correlated with the return in another sector (utility stocks). Specifically, assume that the unconditional expected return on each sector is 12% per year, while the conditional expected return for durables is 10 and 14% following low and high T-bill yield realizations (recessions and expansions), respectively, and the reverse is true for nondurables. A manager who uses the T-bill yield level to vary portfolio weights between 30 and 70% for each sector (averaging a 50% allocation for each) will appear to be informed, according to the unconditional GT measure described in Section 2.1. Specifically, during a transition from a year (quarters 1–4) following a low T-bill yield (a recession) to a year (quarters 5–8) following a high bill yield (an expansion), the manager's 1-year (industry component) GT measure will be measured as:

$$\begin{aligned}
\mathrm{GT}_t &= \sum_{j=1}^{2}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-k-1})\tilde{R}_t^{\mathrm{IND}(j)} \\
&= (\tilde{w}_{\mathrm{durables},4} - \tilde{w}_{\mathrm{durables},0})\tilde{R}_{5-8}^{\mathrm{durables}} + (\tilde{w}_{\mathrm{nondurables},4} - \tilde{w}_{\mathrm{nondurables},0})\tilde{R}_{5-8}^{\mathrm{nondurables}} \\
&= (0.7 - 0.3)(0.14) + (0.3 - 0.7)(0.10) = 0.056 - 0.040 = 0.016 = 1.6\%, \qquad (23)
\end{aligned}$$

where $k = 4$ quarters lag between current and historical (benchmark) portfolios. Thus, the manager will exhibit an outperformance of 160 basis points during the year, based simply on using publicly available information on macroeconomic indicators to shift portfolio weights. Over a large number of periods, this performance measure, in the absence of any private information based skills possessed by the manager, will converge to the covariance between weights and

[23] At the end of each month, the dividend yield is measured as the aggregate cash dividends on the value-weighted CRSP index over the previous 12 months divided by the current level of the index, the default spread is the yield differential between bonds rated BAA by Moodys and bonds rated AAA, and the term spread is the yield differential between Treasury bonds with more than 10 years to maturity and T-bills that mature in 3 months.

returns, which is positive by assumption,

$$\mathrm{GT} = \sum_{j=1}^{2} \mathrm{cov}(\tilde{w}_{j,t-1}, \tilde{R}_{j,t}). \tag{24}$$

To further illustrate, suppose that, in the above example, the probabilities of expansion or recession are each 50% from the point of view of an uninformed observer, but that macroeconomic information gives the manager perfect foresight about whether an expansion or recession will occur during the following year. In this case, the unconditional expected weight, $E[\tilde{w}_{j,t-1}]$, equals 0.5 for each industry, and the GT measure will converge to[24]

$$\begin{aligned}\mathrm{GT} &= \sum_{j=1}^{2} \mathrm{cov}(\tilde{w}_{j,t-1}, \tilde{R}_{j,t}) = \sum_{j=1}^{2} E[(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}])\tilde{R}_{j,t}] \\ &= (2)[0.5(0.7 - 0.5)(0.14) + 0.5(0.3 - 0.5)(0.10)] = 0.008 = 0.8\% \text{ per year.}\end{aligned} \tag{25}$$

If we wish to control for abnormal returns derived from manager weight changes that are correlated with shifts in expected returns (predictable using macroeconomic variable information), then we need to adjust our unconditional measures to take account of such predictability-based returns. We can then make the decision of whether or not we should reward the manager for such public information based returns; the key is to separate abnormal returns based on public information from those derived from private information. Such private information based returns are unambiguously interpreted as manager skill.

3.1. The Ferson–Khang conditional portfolio holdings approach

3.1.1. Description

In response to the need to separate public- and private-based abnormal returns, Ferson and Khang (FK; 2002) develop a conditional portfolio holdings-based performance measurement methodology. FK show that the relation between unconditional and conditional weight-based covariance measures of performance is:

$$\begin{aligned}\underbrace{\sum_{j=1}^{N} \mathrm{cov}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}], \tilde{R}_{j,t})}_{\mathrm{UWM}} &= \underbrace{\sum_{j=1}^{N} E[\mathrm{cov}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}], \tilde{R}_{j,t}|Z)]}_{\mathrm{CWM}} \\ &\quad + \underbrace{\sum_{j=1}^{N} \mathrm{cov}(E[\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}]|Z], E[\tilde{R}_{j,t}|Z])}_{\mathrm{RESID}}.\end{aligned} \tag{26}$$

[24] Note that this is half the level of the performance measure in the prior example, since a recession year is equally likely (by assumption) to be followed by another recession year as it is to be followed by an expansion year. If a recession is followed by a recession, the GT measure will assign zero performance since portfolio weights would not shift from one year to the next for the "perfect foresight" manager.

Here, UWM is the (summed) covariance-based performance measure of Grinblatt and Titman (1993). The right side of Eq. (26) decomposes the GT measure of performance into two components. First, CWM is the performance of the manager, in excess of that attributable to using public information (Z) to choose holdings. The second, RESID, is performance that is due solely to using public information to predict future returns of securities, and varying portfolio weights accordingly. The quality of the manager's private information (skills) is reflected in CWM, while purely mechanical trading strategies are reflected in RESID. While the active portfolio manager should clearly be compensated for producing a positive CWM, similar rewards for producing RESID depend on one's view toward returns that are predictable based on public information variables. An investor should judge the cost of producing such returns, and compare these costs to the fees and other costs charged by the manager who produces returns derived from macroeconomic information that is publicly available.

The CWM is the time-series average of its component at time t, which, in turn, is written as

$$\mathrm{CWM}_t = E\left[\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}])(\tilde{R}_{j,t} - E[\tilde{R}_{j,t}|Z_{t-1}])|Z_{t-1}\right]. \tag{27}$$

An analogous expression captures the time t UWM component,

$$\mathrm{UWM}_t = \sum_{j=1}^{N}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}])(\tilde{R}_{j,t} - E[\tilde{R}_{j,t}]). \tag{28}$$

Note that Eq. (27) involves working with the conditional expected portfolio weight for each security j, $E[\tilde{w}_{j,t-1}|Z_{t-1}]$. In the spirit of Grinblatt and Titman (1993), FK suggest that a good choice of a proxy for these conditional expected weights are the buy-and-hold portfolio weights using the portfolio held k periods ago, and held until the end of period $t-1$, i.e.,

$$\text{proxy for } E[\tilde{w}_{j,t-1}|Z_{t-1}] \equiv \tilde{w}^{b_k}_{j,t-1} = \tilde{w}_{j,t-k-1}\prod_{\tau=t-k}^{t-1}\left(\frac{1+\tilde{R}_{j,\tau}}{1+\tilde{R}_{p,\tau}}\right). \tag{29}$$

However, any reasonable proxy for the expected weights in the presence of conditioning variable information may be used—such as the security weights of a conditionally efficient benchmark portfolio according to some asset-pricing model (e.g., the optimal portfolio of stocks and bonds under the conditional CAPM). The key to choosing a good proxy is that it should reflect the advantage of using current macroeconomic information to improve the portfolio allocation, but not any private information contained in the manager's current (or very recent) weights.

3.1.2. Estimation

FK show that a simple procedure may be used to estimate CWM and UWM. The procedure is as follows:

1. Estimate the (multivariate) regression, $\tilde{R}_{j,t} = \alpha_j + \boldsymbol{b}'_j\boldsymbol{Z}_{t-1} + e_{j,t}$, for each security, j, saving the fitted coefficient vector, $\hat{\boldsymbol{b}}_j$ ($\boldsymbol{Z}_{t-1}$ is the vector of lagged macroeconomic indicators).
2. Estimate the unconditional expected return for each security, j, using the time-series mean return: $\hat{E}[\tilde{R}_{j,t}] = \frac{1}{T}\sum_{t=1}^{T}\tilde{R}_{j,t}$.

3. Using the fitted coefficients from step one above, estimate the regression, $\underbrace{\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-1}^{b_k})(\tilde{R}_{j,t} - \hat{b}_j' Z_{t-1})}_{y\text{-variable}} = \underbrace{CWM}_{\text{intercept}} + \gamma' z_{t-1} + \varepsilon_{j,t}$, to capture the estimated CWM (the intercept), where z_{t-1} is the realization of the vector of macroeconomic variables at the end of time $t-1$, in excess of its long-term (time-series) average.
4. Using the estimated time-series mean return from step two above, estimate UWM as the sample mean of $\sum_{j=1}^{N}(\tilde{w}_{j,t-1} - \tilde{w}_{j,t-1}^{b_k})(\tilde{R}_{j,t} - \hat{E}[\tilde{R}_{j,t}])$.

Estimation of the covariance matrix for the estimated parameters in step 3 above involves simulation methods—the interested reader should refer to FK for details.

3.1.3. Empirical evidence

Ferson and Khang (2002) show that the use of the conditional portfolio weight measure can make a substantial difference in some situations. For their sample of 60 U.S. pension fund managers from 1985 to 1994, Ferson and Khang find that growth fund managers exhibit a positive unconditional weight measure (UWM), but an insignificant conditional weight measure (CWM). They interpret this difference as evidence that growth fund managers exhibit performance simply by smartly using public information to time their stock purchases. As the Ferson and Khang approach is fairly new, it has not yet been applied by other researchers.

4. Conclusion

This paper presented several new methods of performance that take advantage of the additional data embedded in portfolio holdings. In many situations, such data are available, either publicly or privately to a sponsor who is evaluating a manager. We argue that a more precise performance evaluation may be accomplished through the methods outlined in this paper, and that these methods control for some of the basic problems inherent in performance evaluation. The most important of these problems is the choice of a benchmark for a portfolio with evolving (or unknown) style attributes, perhaps due to evolving macroeconomic information.

We recommend the implementation of these approaches when possible, since each has its advantages. With a return-generating model that specifies measurable factors, we recommend the use of the DGTW approach, with the GT approach acting as a check for the accuracy of the benchmarks. In addition, the FK measure allows a control for the influence of macroeconomic variables on security or benchmark returns, when needed. In cases where suitable benchmarks cannot be accurately identified, the GT and FK methods are the preferred approaches.

Acknowledgements

I thank Stan Hurn of Queensland University of Technology for encouraging me to write this survey paper. In addition, I thank the executives and students in my various courses on Performance Evaluation and Attribution (PEVA) worldwide over the past several years for providing useful feedback for my presentation of the notes on which this paper is based. These courses were taught at Chulalongkorn University (Bangkok); the European Central Bank (Frankfurt); the FAME Executive Education program (Geneva); Queensland University of Technology (Brisbane); Stockholm University; the Quantitative Finance Research Center at the University of Technology, Sydney; and the University of Vienna.

References

Brinson, G., Hood, L., & Beebower, G. (1986). Determinants of portfolio performance. *Financial Analyst Journal*, *42*, 39–44.

Avramov, D., & Wermers, R. (2006). Investing in mutual funds when returns are predictable. *Journal of Financial Economics*, forthcoming.

Brown, S., & Goetzmann, W. (1995). Performance persistence. *Journal of Finance*, *50*, 679–698.

Carhart, M. (1997). On persistence in mutual fund performance. *Journal of Finance*, *52*, 57–82.

Chen, H., Jegadeesh, N., & Wermers, R. (2000). The value of active mutual fund management: An examination of the stockholdings and trades of fund managers. *Journal of Financial and Quantitative Analysis*, *35*, 343–368.

Copeland, T., & Mayers, D. (1982). The value line enigma (1965–1978): A case study of performance evaluation issues. *Journal of Financial Economics*, *10*, 289–322.

Daniel, K., Grinblatt, M., Titman, S., & Wermers, R. (1997). Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance*, *52*, 1035–1058.

Daniel, K., & Titman, S. (1997). Evidence on the characteristics of cross-sectional variation in common stock returns. *Journal of Finance*, *52*, 1–33.

Datar, V., Naik, N., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets*, *1*, 203–219.

Fama, E., & French, K. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, *19*, 3–29.

Fama, E., & French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, *47*, 427–465.

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–56.

Fama, E., & French, K. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, *51*, 55–84.

Ferson, W., & Harvey, C. (1999). Conditioning variables and the cross section of stock returns. *Journal of Finance*, *54*, 1325–1360.

Ferson, W., & Khang, K. (2002). Conditional performance measurement using portfolio weights: Evidence for pension funds. *Journal of Financial Economics*, *65*, 249–282.

Ferson, W., & Schadt, R. (1996). Measuring fund strategy and performance in changing economic conditions. *Journal of Finance*, *51*, 425–461.

Grinblatt, M., & Titman, S. (1989a). Mutual fund performance: An analysis of quarterly portfolio holdings. *Journal of Business*, *62*, 394–415.

Grinblatt, M., & Titman, S. (1989b). Portfolio performance evaluation: old issues and new insights. *Review of Financial Studies*, *2*, 393–422.

Grinblatt, M., & Titman, S. (1992). The persistence of mutual fund performance. *Journal of Finance*, *47*, 1977–1984.

Grinblatt, M., & Titman, S. (1993). Performance measurement without benchmarks: An examination of mutual fund returns. *Journal of Business*, *66*, 47–68.

Grinblatt, M., Titman, S., & Wermers, R. (1995). Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *American Economic Review*, *85*, 1088–1105.

Grossman, S., & Stiglitz, J. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, *70*, 393–408.

Gruber, M. (1996). Presidential address: Another puzzle: The growth in actively managed mutual funds. *Journal of Finance*, *51*, 783–810.

Hendricks, D., Patel, J., & Zeckhauser, R. (1993). Hot hands in mutual funds: Short-run persistence of relative performance (1974–1988). *Journal of Finance*, *48*, 93–130.

Iihara, Y., Kato, H., & Tokunaga, R. (2004). The winner-loser effect in Japanese stock returns. *Japan and the World Economy*, *16*, 471–485.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance*, *45*, 881–898.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, *48*, 65–91.

Jeng, L., Metrick, A., & Zeckhauser, R. (2003). Estimating the returns to insider trading: A performance-evaluation perspective. *Review of Economics and Statistics*, *85*, 453–471.

Jensen, M. (1968). The performance of mutual funds in the period 1945–1964. *Journal of Finance*, *23*, 389–416.

Jern, B. (2002). Can past performance be used to find the winners among international stock funds? *Ekonomiska Samfundets Tidskrift*, *55*, 77–97.

Kacperczyk, M., Sialm, C., & Zheng, L. (2005). On the industry concentration of actively managed equity mutual funds. *Journal of Finance*, *60*, 1983–2012.

Kandel, E., Kandel, S., & Wermers, R. (2005). Endogenous benchmarks. Working paper.

Keim, D., & Stambaugh, R. (1986). Predicting returns in the stock and the bond markets. *Journal of Financial Economics*, *17*, 357–390.

Lee, C., & Swaminathan, B. (2000). Price momentum and trading volume. *Journal of Finance*, *55*, 2017–2069.

Moskowitz, T., & Grinblatt, M. (1999). Do industries explain momentum? *Journal of Finance*, *54*, 1249–1290.

Nitibhon, C., Tirapat, S., & Wermers, R. (2005). Secrets of Thai equity funds revealed: An analysis of performance, persistence, and flows. Working Paper.

Pinnuck, M. (2003). An examination of the performance of the trades and stock holdings of fund managers: Further evidence. *Journal of Financial and Quantitative Analysis*, *38*, 811–828.

Roll, R. (1978). Ambiguity when performance is measured by the securities line. *Journal of Finance*, *33*, 1051–1069.

Roll, R. (1992). A mean/variance analysis of tracking error. *Journal of Portfolio Management*, *18*, 13–23.

Rouwenhorst, G. (1998). International momentum strategies. *Journal of Finance*, *53*, 267–284.

Wermers, R. (1997). Momentum investment strategies of mutual funds, performance persistence, and survivorship bias. Working Paper.

Wermers, R. (1999). Mutual fund herding and the impact on stock prices. *Journal of Finance*, *54*, 581–622.

Wermers, R. (2000). Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance*, *55*, 1655–1695.

Wermers, R. (2002). A matter of style: The causes and consequences of style drift in institutional portfolios. Working Paper.

Wermers, R. (2005). Is money really smart? New evidence on the relation between mutual fund flows, manager behavior, and performance persistence. Working Paper.

Wylie, S. (2005). Fund manager herding: A test of the accuracy of empirical results using U.K. data. *Journal of Business*, *78*, 381–403.