Benchmarking Money Manager Performance: Issues and Evidence

Louis K. C. Chan
University of Illinois at Urbana-Champaign

Stephen G. Dimmock
Michigan State University

Josef Lakonishok
University of Illinois at Urbana-Champaign

Academic and practitioner research evaluates portfolio performance using size and value/growth attributes or factors. We assess the merits of popular evaluation procedures based on matched-characteristic benchmark portfolios or time-series return regressions by applying them to a sample of active money managers and passive indexes. Estimated abnormal returns display large variation across approaches. The benchmarks typically used in academic research—attribute-matched portfolios from independent sorts, the three-factor time-series model, and cross-sectional regressions of returns on stock characteristics—track returns poorly. Some simple alterations improve the performance of these methods. (*JEL* G11, G12, G14, G23)

Active money managers offer the allure of returns that exceed market benchmarks. Managers with successful track records are hotly pursued by investors, while those who fall short of their targets are eventually fired. Investors' close scrutiny of a portfolio manager's performance highlights the importance of establishing relevant benchmarks. The research literature provides many procedures. Earlier studies such as Jensen (1968) use the capital asset pricing model (CAPM) to generate expected returns. More recent work draws on Chan, Hamao, and Lakonishok (1991), Fama and French (1992), and Lakonishok, Shleifer, and Vishny (1994), who find that size, and the ratio of book-to-market value of equity, capture much of the variation in returns across stocks. The use of these two attributes in performance evaluation is now pervasive in academic research. Performance is measured either through regressing

We thank Kent Daniel, Eugene Fama, Ravi Jagannathan, Jason Karceski, Matthew Spiegel (the editor), Bhaskaran Swaminathan, two anonymous referees, seminar participants at the Hong Kong University of Science and Technology, Kellogg Hedge Fund Conference at Northwestern University, National University of Singapore, and the University of Texas at Austin for comments, and John Diderich, James Owens, Menno Vermeulen, and Simon Zhang for assistance with data. Send correspondence to Louis K. C. Chan, 340 Wohlers, 1206 South Sixth Street, Champaign, IL 61820; telephone: 217-3336391. E-mail: l-chan2@uiuc.edu.

doi:10.1093/rfs/hhp016 Advance Access publication April 13, 2009

a manager's returns on the returns of portfolios that mimic the market, size, and book-to-market factors as in Fama and French (1993), or by comparing returns to those on a passive portfolio of stocks with similar size and book-to-market characteristics.

In practice, many investment consultants draw on academic research to develop benchmarks for performance evaluation and attribution. Managers' performance relative to such indexes guides the allocation of investment mandates worth billions of dollars. Some of the earlier yardsticks, for instance the Standard & Poor's BARRA indexes until 2005, parallel academic studies in terms of using size and book-to-market as the sole attributes for stock classification. Other indexes consider additional variables, such as analysts' long-term growth forecasts in the case of the Russell indexes. More recently there has been a trend in the industry toward customized benchmarks to adjust for investment style along the dimensions of size and value/growth orientation (see Chan, Chen, and Lakonishok 2002 and Ben Dor, Jagannathan, and Meier 2003). By identifying a manager's style, the active portfolio can be paired with a passive benchmark that mimics the underlying strategy. As a result, stock selection skills may be detected more clearly.

The upshot is that academic research and industry practice yield a proliferation of methods using size and value/growth attributes or factors as the basis for evaluating portfolio performance. At first glance, because they are variants of the same underlying approach, all these methods are perceived to be more or less interchangeable. For example, Fama and French (1992) find that in the cross-section, the effect of stocks' earnings-to-price ratios is absorbed by size and book-to-market. Additionally, the three-factor model in Fama and French (1996) captures the returns on portfolios sorted by ratios of earnings or cash flow to price, or sorted by sales growth. A casual interpretation of these results is that other indicators of a portfolio's value/growth orientation are unimportant once book-to-market is accounted for. Similarly, on the surface it may appear that cross-sectional regression methods and time-series factor models yield similar conclusions with respect to detecting abnormal returns. Perhaps on the basis of this evidence, Fama and French (1993) say that "evaluating the performance of a managed portfolio is straightforward" using their three-factor model.

Such complacency may be unwarranted. Rather, the evidence indicates that variants of the size and value/growth benchmarking procedure produce serious disagreements about the existence and level of abnormal returns. To give a foretaste of our findings, we use two procedures that are standard in the academic literature to evaluate the performance of a sample of 199 institutional money managers (the full details of the sample and benchmarking procedures are presented in the following sections). A managed portfolio's performance is measured as its quarterly return in excess of a benchmark. In the first procedure, the yardstick is a reference portfolio that matches the size and book-to-market characteristics of each stock in the managed portfolio. We use

25 reference portfolios produced from independent sorts on size and book-to-market. In the second procedure, the benchmark return is the fitted value from the Fama-French (1993) three-factor model time-series regression applied to the entire return history of the managed portfolio.

If the procedures are closely aligned, for a given portfolio they should deliver average abnormal returns that are of the same sign (over- or underperformance). Over the full sample period of 1989–2001, the methods disagree on the sign of excess return in about one out of four portfolios (24.62% of the cases). As further cause for concern, the mean annualized abnormal returns frequently diverge by large magnitudes. For the overall sample, the levels of the absolute differences exceed 2.5% in 43.22% of the portfolios and are at least 5% in 14.07% of the cases. During the volatile period from 1998 to 2000, the differences are much more pronounced. For instance, in this subperiod, absolute differences above 5% occur for 43.75% of the portfolios.

The message from these results is that two seemingly interchangeable benchmarking procedures can produce very different results with economically important consequences. Other, potentially superior, methods for deriving reference portfolios are widely used as well, so these comparisons are only illustrative. In this paper, we explore different benchmarking procedures as they are implemented in research and practice. The goal is to gauge how the choice of benchmarking procedure affects inferences about investment performance, trace the underlying sources of the differences, identify any potential shortcomings in the procedures, and suggest improvements. In particular, our discussion focuses on three key choices in benchmark construction, and their consequences for measured performance. First, we examine the use of independent sorts to determine the size and book-to-market control portfolios. Second, we put the time-series three-factor model with the market and zero-investment mimicking portfolios up against style-based return regressions using the full assortment of equity asset classes. Lastly, we analyze how well a portfolio's value/growth orientation is captured by looking only at book-to-market.

We apply the benchmarking procedures to two sets of data. To ensure that our test environment captures all the conditions that would exist in a typical evaluation or attribution exercise, we examine a sample of large institutional money managers over the 1989–2001 period. We provide comparisons of methods averaged over managers, and comparisons of how individual managers are ranked. Additionally, we take the methods to the returns on widely used passive indexes, whose composition follows clearly prespecified criteria.

The evidence in this paper has implications beyond the evaluation of managed portfolios' performance. Any analysis of long-term stock price performance invariably grapples with the choice of an appropriate benchmark. The issue is central in studies of stock market efficiency, such as tests of the profitability of trading strategies. Research on the impact of various managerial decisions, such as equity offerings, dividend initiations or omissions, and share repurchase programs, also faces the problem of measuring stock returns in excess of some

normal level. A few other studies, such as Barber and Lyon (1997); Fama (1998); Lyon, Barber, and Tsai (1999); and Mitchell and Stafford (2000), alert researchers to the hazards of testing for abnormal returns in multi-year event studies.

Our key result is that judgments about the magnitude of performance are sensitive to the benchmarking methodology. To illustrate, mean abnormal returns are 2.64% relative to the Fama-French three-factor model, 1.39% when compared to reference portfolios based on independent sorts on size and book-to-market, a measly 0.78% when we use a more comprehensive measure of value/growth, and drop to −1.97% relative to a benchmark from cross-sectional regressions of returns on stock attributes. These differences stand out all the more because they are averages across an extensive sample of portfolios over many quarters. Inferences about performance are fragile even though all our procedures rest on the same basic premise that a portfolio's size and value/growth orientation are central determinants of its expected return. A true skeptic might conclude that the root of the problem lies with the general three-factor model's inability to capture well the behavior of returns, regardless of how the factors are approximated.

Tracking error volatilities provide a way to judge how well the benchmarks capture the behavior of active portfolios. In this respect, benchmarks from procedures that are widely used in academic research disappoint, yielding high tracking error variability. We trace their relatively poor showing to the underlying drawbacks—independently sorting stocks by size and book-to-market, treating the effects of size and value/growth as linear additive terms that are uniform across all stocks, and relying on book-to-market as the sole yardstick for value/growth classification. Conversely, methods that bypass these shortcomings fare better. In particular, attribute-matched benchmark portfolios formed from sorts on size and then on a comprehensive indicator of value/growth orientation within each size category produce relatively low tracking error volatility. Out-of-sample tracking error volatilities average 10.54% under the conventional three-factor model, while dollar-weighted reference portfolios that match the size- and composite value characteristics of active managers deliver mean volatilities of 8.71%. More generally, evidence from the passive Russell indexes indicates that the characteristic-matched benchmarking procedures have better tracking ability than the regression-based procedures.

The remainder of the paper is organized as follows. Section 1 describes our data and outlines some key issues regarding benchmark construction. To streamline the discussion, we discuss separately benchmarking procedures based on portfolio holdings and those based on return regressions. Section 2 provides results on investment performance based on characteristic-matched baseline portfolios. To explore further the sources of the differences across benchmarking procedures, Section 3 applies them to passive portfolios as given by the Russell style indexes, and provides details on the characteristics of

the benchmark portfolios. Results on money manager performance relative to regression-based benchmarks are provided in Section 4. Some diagnostics on how the regression-based benchmarks fare, including their performance on passive indexes, are contained in Section 5. Section 6 examines how the results for a managed portfolio vary with the choice of benchmarking procedure. Section 7 provides concluding remarks.

1. Preliminaries

1.1 Data

Our sample describes the quarterly returns and holdings from 1989:Q1–2001:Q4 of 199 U.S. institutional equity portfolios offered by investment management firms. The firms include many of the largest and most prominent money managers in the industry. At year-end 2001, for example, their assets under management amounted to about $5 trillion. The portfolios span a variety of styles in terms of size and value/growth orientation. While the portfolios vary in terms of when their return histories start and end, we require that each has at least 16 consecutive quarters of returns. The data are collected by SEI Investments, a large investment services firm.

The dataset is not free of selection bias, as larger, relatively more successful managers are more likely to be included in the database. Nonetheless, it is representative of performance databases maintained by investment consulting firms that are widely used in clients' searches for portfolio managers.

1.2 Approaches to performance evaluation

Although we concentrate on the evaluation methods that are most widely used in research and practice, the list is by no means exhaustive. One alternative follows Jegadeesh and Titman (1993), who find that stock returns over intermediate horizons are related to lagged returns. Daniel et al. (1997), for example, sort by size, book-to-market, and past 12-month return to build reference portfolios. Our analysis does not incorporate the momentum factor for the following reasons. Most notably, the investment community has not viewed momentum as a distinct style. Rather, institutional clients hold their managers to passive yardsticks based on size and value/growth. This is reflected by the battery of style indexes, such as those by Russell, MSCI/BARRA, and Dow Jones Wilshire, that are extensively used in the industry. Clients and consultants, on the other hand, do not specify yardsticks directly based on past return. Because we want to correspond as much as possible to evaluation methods that are applied in practice, our benchmarks do not consider explicitly the role of past returns. In any event, it is not clear *a priori* whether the omission of momentum effects systematically discriminates in favor of one evaluation approach over another. A final reason for our choice to overlook the momentum effects is tractability. An entire set of issues surrounds momentum-based reference portfolios, such

as the choice between independent or conditional sorts and the specification of the horizon for past return. To keep the discussion to a manageable length, therefore, we only concentrate on the key dimensions of size and value/growth.

Other approaches to performance evaluation draw on Admati and Ross (1985) and Dybvig and Ross (1985), who show that the portfolio of an investor with superior private information earns a return that is nonlinearly related to benchmark returns. Merton (1981) and Glosten and Jagannathan (1994) suggest that the nonlinear component can be viewed as outcomes of various option strategies. Goetzmann et al. (2007) develop a performance ranking metric that is robust to nonlinear payoffs from managed portfolios. Ferson and Schadt (1996) and Ferson and Khang (2002) argue in favor of using conditioning information to capture time variation in risks and returns, or the effects of dynamic investment strategies. The investment industry has not widely adopted such techniques, however (at least for managers limited to long positions in equities). Moreover, their implementation raises many questions on how to model the contingent claim aspect of the nonlinear payoffs or the relationship between conditioning information and returns, the appropriate empirical proxies for the option strategies, and the selection of instruments. Again for the sake of tractability, we do not consider examples of these alternative evaluation methods.

1.3 Issues in benchmark construction

Even in the context of the general three-factor approach, there is a wide variety of ways to construct benchmark returns. In broad terms, the choices involve the use of stock attributes or loadings from regression models; the specific measures of value/growth orientation; whether size and value/growth are treated independently; the weighting scheme for stocks in the benchmark; and the frequency with which the benchmark's composition is updated.

1.3.1 Attributes or loadings. Daniel and Titman (1997) find that stock attributes do a better job than factor loadings in predicting the cross-section of returns. Accordingly, one approach is to obtain benchmark returns from attribute-sorted portfolios that match the features of the stocks held by the active manager. Each holding in the active portfolio is paired with a reference portfolio that mimics as closely as possible the stock's size and value/growth tilt. The weighted average of the matching portfolios' returns over all holdings yields the benchmark return for the active portfolio. Daniel et al. (1997) and Wermers (2004) apply this "characteristic-based" approach to study the performance of U.S. equity mutual funds. Instead of using reference portfolios, the return can be predicted from a cross-sectional regression of stock returns on beginning-of-quarter stock attributes.[1]

[1] The BARRA performance attribution system, which is heavily used in the investment industry, is based on such a cross-sectional regression approach.

However, timely data on managers' portfolio holdings are not generally available. Many studies therefore estimate expected returns with factor loadings from time-series regressions of portfolio returns on proxies for the factors. Carhart (1997) is one example of this "regression-based" approach to performance measurement.

1.3.2 Measuring value/growth style. In many studies, a stock is considered as value or growth solely on the basis of its book-to-market ratio.[2] Similarly, factor loadings with respect to a zero-investment mimicking portfolio that is long (short) in stocks with high (low) book-to-market ratios are used to assign stocks to value or growth categories. As Lakonishok, Shleifer, and Vishny (1994) note, however, the ratio of book-to-market value of equity is an incomplete measure of a stock's value/growth orientation. For example, under current U.S. accounting standards, book values do not include the value of intangible capital, such as investments in research and development (see Chan, Lakonishok, and Sougiannis 2001). Similarly, until recently, measured book values ignored the underfunding of companies' pension liabilities. Looking at other indicators, such as earnings, dividends, or sales, may help to paint a clearer picture of a stock's value/growth stance. As an illustration, on the basis of book-to-market ratios, pharmaceutical companies during the late 1990s would have been classified as similar to Internet-oriented companies. This would have glossed over the important difference that pharmaceutical companies generally had a proven past record of sales and profitability, unlike Internet-oriented firms.

1.3.3 Independence of size and value/growth classification. Reference portfolios used in research and practice are generally formed from two-way sorts on size and book-to-market equity. A crucial issue is whether the sorts are done independently, or within a particular group. In one-way sorts by book-to-market, the growth (low book-to-market) category tends to comprise larger stocks than the value (high book-to-market) category. Intersecting this classification with an independent sort by size thus results in large stocks generally being clustered in the growth category. The problem is that this classification provides a poor depiction of money managers' investment domains. Many investment managers tend to concentrate on larger stocks, where information as well as liquidity tends to be more available. Within the category of large stocks, some managers, who are more value-oriented, seek out comparatively cheap, undervalued stocks that have attractive earnings or dividend yields. Other large-capitalization managers who are more glamour-oriented focus on stocks with high growth potential, or substantial investments in intangible capital. Despite the differences in their approaches, an independent classification scheme might

[2] The academic research literature generally has not addressed issues related to the measurement of size. While market capitalization is one choice, adjustments for cross-holdings or privately held shares present other possibilities.

hold both groups of managers to similar benchmarks (large stocks with low book-to-market ratios).

An alternative to the classification scheme based on independent sorts is to define value and growth within each size category. This corresponds more closely to how portfolio managers structure their stock selection process, whereby a manager may choose, for example, relatively cheaper stocks within mid-sized firms. As evidence of the pervasiveness of this practice, many widely used market indexes, such as those produced by the Frank Russell Company, S&P Citigroup, and Wilshire Associates, define value or growth within groups of similarly-sized firms.

1.3.4 Choice of basis portfolios. Studies adopting the regression-based approach to performance measurement typically use as proxies the returns on basis portfolios that are highly correlated with the factors. Identifying these basis portfolios is not a clear-cut matter, however.[3] Fama and French (1993) use a market index as well as zero-investment portfolios from independent sorts on size and book-to-market. Ben Dor, Jagannathan, and Meier (2003), following Sharpe (1992), use the returns on a full set of equity style indexes. Basis portfolios that do poorly in capturing the behavior of the factors will induce large errors in judgments about performance. Moreover, some proxies that mimic the variation in the factors may not correspond well to the investment opportunities relevant to active managers. As a result, active portfolios may have negligible loadings on such benchmarks and measured exposures will not align with the true exposures. Yardsticks based on these reference portfolios will thus be associated with large benchmark errors.

1.3.5 Weighting scheme. A benchmark is intended to capture the performance of a representative set of stocks that share similar features. It is thus undesirable that the benchmark's behavior is driven by a relatively small subset of the underlying stocks. Equally weighting the component stocks prevents the behavior of the yardstick from being dominated by idiosyncratic shocks to a few companies. However, this tends to give relatively more weight to smaller stocks in the benchmark. Value-weighting the component stocks, on the other hand, tends to emphasize larger stocks whose returns are generally less noisy. Further, value-weighting mitigates biases in computing expected returns induced by rebalancing.

1.3.6 Frequency of reconstitution. A stock's attributes may change over time, so a reference portfolio that originally represents stocks with similar features may become less homogeneous. The ability of the reference portfolio to track the active portfolio's return may thus deteriorate over time. Reconstituting

[3] Lehmann and Modest (1987) discuss some of the issues involved in constructing the basis portfolios in the context of the Arbitrage Pricing Theory.

the reference portfolio more frequently alleviates the problem. Suppliers of benchmark indexes, for example, update their indexes every quarter (Wilshire) or once a year (Russell).

Since our collective understanding of the return-generating process is incomplete, it is important to ensure that the performance results do not hinge upon the choice of a benchmarking model. Accordingly, in our evaluation of money manager performance in the subsequent sections, we employ assorted methods representing different choices with respect to each of the considerations above.

2. Performance Relative to Characteristic-Matched Portfolios

We discuss performance relative to characteristic-matched benchmarks using portfolio holdings in this section. The analysis of performance under regression-based benchmarks is deferred to the next section.

2.1 Methods

We use four versions of characteristic-matched reference portfolios. In every case, the benchmark for a given manager is constructed as follows. Each stock in the managed portfolio, based on its size and value/growth attribute ranks, is paired with a reference portfolio. The benchmark return is then the weighted average of the buy-and-hold quarterly returns of the control portfolios, using the manager's investment weights as of the beginning of the quarter.

2.1.1 Independent size, book-to-market sorts. In the first procedure, we use independent sorts to form reference portfolios for each size and value/growth category. This procedure mirrors Fama and French (1993) and subsequent studies including Lakonishok and Lee (2001) and Chan, Lakonishok, and Sougiannis (2001). The control portfolios are formed once a year in July. The sort on size (the market value of common equity of the stock as of the end of June) yields five portfolios based on NYSE breakpoints. Independently, stocks are ranked and sorted into quintile portfolios by the ratio of book to market value of common equity (also based on NYSE breakpoints). Book value is from the latest fiscal year ending in the prior calendar year, while market value is from December of the previous year-end. There are 25 control portfolios from the intersection of these two sorts. The return on each portfolio is either the equally weighted or value-weighted average of the buy-and-hold returns on the component stocks.

2.1.2 Size-conditional book-to-market sorts. Ikenberry, Lakonishok, and Vermaelen (1995) and Daniel et al. (1997) partition stocks into value and growth categories for similarly-sized firms. We implement a size-conditional classification as follows. At the end of June each year, we define six categories of firms by market value of equity, moving from the largest to the smallest stock in the listed U.S. domestic common equity universe. The categories are

set up so that each represents a meaningful share of total capitalization, while still comprising a fairly large number of firms. The first group is made up of the top 75 stocks by market value, while the second includes the next 125 largest, then the next 300 largest make up the third, the following 500 stocks are placed in the fourth, the next 1000 stocks in order of size are in the fifth, and the remainder make up the last group.[4] Within each size category, we rank stocks by the ratio of book value of equity (as of the prior fiscal year) to market value of equity (as of December in the prior year) and classify them from relatively value-oriented to relatively growth-oriented (with high and low book-to-market ratios, respectively). Since the first category by firm size (the largest 75 stocks) contains a relatively small number of stocks, it is divided into only three groups by value/growth (with an equal number of stocks in each group); in each of the other size classifications, there are five groups by value/growth, with roughly equal numbers of stocks.[5] This size-conditional book-to-market classification scheme yields a total of 28 portfolios. Within each portfolio, the buy-and-hold returns on the component stocks are either equally weighted or value-weighted.

2.1.3 Size-conditional composite value/growth indicator approach. Our third approach does not rely solely on book-to-market as the indicator of value/growth; moreover, a stock's value/growth orientation is defined relative to similarly-sized firms. Specifically, we construct a composite indicator variable to measure value/growth orientation. The composite is the rescaled average of a stock's percentile rank on each of five attributes, such that the most value-oriented (growth-oriented) stock within a given size category receives a rank of one (zero). The five characteristics are book-to-market ratio, sales-to-price ratio, cash flow to firm value, dividend yield, and earnings yield. Once ranks are calculated over as many of these variables as are available, the simple average is computed.[6] Stocks within a given size category are ordered from lowest to highest by this average, which is then rescaled to range from zero to one. The return on the reference portfolio is either an equally weighted or value-weighted average of the underlying stocks' returns.

---

[4] The largest 75 stocks make up on average 45% of total equity market capitalization, while the other groups represent on average 20%, 15%, 10%, 6%, and 4%, respectively.

[5] Alternatively, as is the case with many indexes used in the investment community, each size class can be separated into value/growth subsets with roughly equal market capitalization. To provide a more direct comparison with benchmarking methods used in academic research, we do not follow this approach.

[6] Characteristics are calculated in July each year. Values for accounting variables are taken from the latest fiscal year as of the prior year-end, and are scaled by stock price or market capitalization in December of the previous calendar year. Sales-to-price is net sales divided by equity market capitalization. Cash flow to firm value is operating income before depreciation divided by firm value (total assets less book value of common equity, minus accounts payable, plus market value of common equity). Dividend yield is cash dividends to common equity divided by equity market capitalization. Earnings yield is income before extraordinary items available to common equity divided by equity market capitalization. Negative values for the accounting variables are treated as follows. As in Fama and French (1993), stocks with negative book values of equity are excluded from the analysis. Cases with negative values for net sales, cash flow, or earnings, and firms not paying dividends, are assigned ranks of zero for the respective variable. The remaining cases with nonnegative values or positive dividends are ranked from lowest to highest.

2.1.4 Quarterly size-conditional book-to-market sorts. In the above three procedures, the control portfolios are updated once a year at the end of June. Using stale data may mean that a reference portfolio's underlying characteristics (hence its expected return) are not fully aligned with the active portfolio. To achieve a closer correspondence, our fourth procedure uses size-conditional book-to-market matched portfolios where the control's composition is updated at the end of every quarter using current quarter-end market capitalization. As with the other methods, we report both equally weighted and value-weighted returns.

2.1.5 Russell style indexes. Finally, as a baseline comparison for our reference portfolios, we use the Russell style indexes. In practice, these are the most commonly used benchmarks for institutional equity investors. In the majority of cases, the money manager reports the portfolio's style, and we assign to each active portfolio the Russell index corresponding to its style. When the style description is unavailable, we use the portfolio's ranks by size and composite value indicator to determine its style. Specifically, a portfolio's weighted average size percentile rank (one for the largest stock and zero for the smallest stock) across its holdings determines the manager's size orientation. Size ranks above 0.8 are classified as large; size ranks between 0.8 and 0.6 are classified as mid-cap; size ranks below 0.6 are treated as small. A manager's composite value score (one for the most value-oriented and zero for the most growth-oriented) determines the value/growth orientation. Indicator values above 0.67 denote value; those below 0.33 denote growth; the intermediate range is classified as "neutral."[7] Large, mid-cap, and small capitalization value or growth managers are paired with the appropriate Russell 1000, Russell mid-cap, and Russell 2000 value or growth index. Neutral portfolios are compared against the Russell core index for the corresponding size category.

2.2 Results

In accord with standard practice in the investment management industry, a portfolio's average abnormal return is its time-series geometric mean annual return minus the time-series geometric mean annual return on the matched benchmark. In the absence of any stock selection ability, the average abnormal return should be close to zero. While a reference portfolio may be unbiased in the sense that on average it yields the same return as an active portfolio, it may nonetheless fail to track the managed portfolio's return closely. As a result, the control procedure may yield unreliable inferences about performance. Everything else being equal, a benchmark that tracks better the active portfolio raises confidence that a manager's differential performance is due to skill rather than luck. Accordingly, we also examine tracking error volatility under each

[7] As a check on our choice of cutoffs, we use them to confirm the managers' self-declared styles when they are reported and find that they generally agree. Chan, Chen, and Lakonishok (2002) find that mutual fund portfolios' attributes provide good guidance on their investment styles.

of the methods, defined as the annualized standard deviation of the quarterly differences between the portfolio's return and the benchmark's return. To the extent that the benchmark portfolio matches the manager's investment domain, the tracking error volatility should be low.

Table 1 summarizes the distribution of abnormal return and tracking error volatility across the sample of money managers for each method. The cross-sectional average and median are reported for the entire sample period and also for the 1998:Q1–2000:Q1 subperiod.

The procedures in Table 1 share the same underlying viewpoint about what drives stock returns. Moreover, the results are averaged across a broad sample of portfolios over many years. The presumption therefore is that any differences across the methods should be meager. The results reveal, however, that the level of excess returns varies markedly across methods. Abnormal returns range from 2.72% to 0.71% for the overall sample period, yielding a range of 2.01%. Median abnormal returns display a similar range. Put another way, what might appear to be slight variations of the same underlying methodological approach translate into quite different conclusions about the level of performance.

Notably, the largest abnormal return, 2.72%, arises when the Russell indexes are used as the benchmark. The equally weighted portfolio of managers earns a mean quarterly abnormal return that is 4.90 standard errors away from zero.[8] Hence, generic indexes that have a wide following in the investment industry suggest reliably high levels of performance for our set of managers. This finding probably reflects the selection bias underlying the sample: many databases that are used in the investment industry to track performance, such as the one we use here, are designed to aid in selecting superior managers. Since performance in practice is usually measured against the Russell benchmarks, those managers who stand out against them are more likely to be included.

Tracking error volatility indicates how closely a benchmark return series covaries with active portfolio returns. Reference portfolios from independent sorts yield the highest tracking error volatilities on average. Equally weighted benchmarks under this procedure generate a mean tracking volatility of 10.37% per year. Given the lower variability in the returns on large stocks and their stronger covariation, the tracking error volatility is reduced to 9.35% when the control portfolios are value-weighted. In comparison, when we use a finer size classification and measure book-to-market ranks within similarly-sized firms, volatilities drop to 9.51% (8.97%) for the equally weighted (value-weighted)

[8] Insofar as managers follow similar strategies and pick some of the same stocks, abnormal returns are cross-sectionally correlated. Significance tests based on the cross-sectional standard deviation are thus misleading. To avoid this problem, we calculate the abnormal return on an equally weighted portfolio of all money managers in the sample in the quarter based on a given method. The time-series volatility of this return builds in the cross-sectional correlation, and lets us check whether in the time series the mean is significantly different from zero. When we test for the equality of mean abnormal returns across the procedures in Table 1 for the equally weighted portfolio of managers, the F-statistic (with standard errors adjusted for clustering by time) is 2.89 with a p-value of 0.01.

Table 1
Performance (in percentage per year) of managed portfolios using alternative characteristics-based benchmarks

		Annual independent size, BM		Annual size, within-size, BM		Annual size, value composite		Quarterly size, within-size, BM		
		Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight	Russell index
Panel A: Full period, 1989:Q1–2001:Q4										
	Mean	2.06	1.39	1.33	1.13	0.79	0.78	0.72	0.71	2.72
Abnormal return	Median	1.96	1.10	0.94	0.87	0.60	0.45	0.46	0.33	2.26
	Std Dev	4.26	4.27	4.50	4.63	4.41	4.47	4.57	4.64	4.50
	t-stat	1.38	3.57	1.52	3.02	1.33	1.46	1.72	2.23	4.90
Tracking error	Mean	10.37	9.35	9.51	8.97	8.72	8.71	9.01	8.80	8.94
volatility	Median	9.00	7.90	8.16	7.80	7.28	6.86	7.74	7.46	7.92
Mean absolute	Size	0.028	0.010	0.029	0.008	0.034	0.014	0.035	0.015	0.093
difference of	Book-to-market	0.031	0.031	0.012	0.010	0.029	0.029	0.012	0.010	0.067
characteristic ranks	Value composite	0.089	0.082	0.064	0.059	0.017	0.012	0.065	0.060	0.090
Panel B: 1998:Q1–2000:Q1										
	Mean	3.37	0.65	3.14	0.70	2.12	1.67	0.75	0.84	2.21
Abnormal return	Median	1.09	−2.31	1.06	−1.19	1.28	0.77	−1.53	−2.00	0.54
	Std Dev	14.08	14.13	13.55	13.20	10.85	11.21	13.64	13.94	12.59
Tracking error	Mean	10.70	10.29	10.04	9.98	9.43	9.44	10.02	9.81	9.34
volatility	Median	8.80	8.28	7.82	7.90	7.32	7.30	7.86	8.08	7.52
Mean absolute	Size	0.027	0.011	0.029	0.008	0.034	0.013	0.036	0.015	0.090
difference of	Book-to-market	0.031	0.032	0.012	0.010	0.028	0.028	0.012	0.010	0.072
characteristic ranks	Value composite	0.092	0.087	0.066	0.062	0.017	0.012	0.068	0.063	0.089

At the beginning of a quarter, every stock held in a managed portfolio is matched with a control portfolio based on its characteristics, using one of several procedures. The benchmark return is the weighted average of the quarterly buy-and-hold returns on the control portfolios. The procedure is repeated every quarter. A managed portfolio's performance is measured as its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4), and during 1998:Q1–2000:Q1. A portfolio's mean abnormal return is its annualized geometric mean return minus the annualized geometric mean return on the benchmark. A portfolio's tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the portfolio's return and the benchmark's return. For each performance measure, the arithmetic mean and median are provided over the cross-section of 199 managed portfolios in the sample period. The reported *t*-statistic is for the hypothesis that the time-series mean of the quarterly equally weighted average excess return over the benchmark across all available managed portfolios equals zero. Also reported are mean absolute differences between the characteristic ranks of the managed portfolios and benchmarks under each procedure with respect to size, book-to-market, and the composite value indicator. The procedures to construct control portfolios are as follows. Under the independent sorting procedure, there are 25 control portfolios from the intersection of independent sorts by size (market value of equity) and BM (book-to-market ratio: the ratio of book value of common equity to the market value of common equity). Under the size, within-size BM sort procedure, there are 28 control portfolios from sorts first by size, and then within each size category, by BM. In the size, value composite approach, a stock is given an overall ranking, conditional on its size group, based on book-to-market, dividend yield, cash flow yield, average earnings yield (based on the past year's net income, forecasted next year earnings, and forecasted two-year ahead earnings), and sales-to-price ratio. In these methods, the component stocks in a control portfolio are refreshed once a year at the end of June. In the quarterly size, within-size BM approach, the component stocks are refreshed at the beginning of each quarter. The return on a control portfolio is either the equally weighted or value-weighted average of the buy-and-hold returns on the component stocks. Each managed portfolio is also paired with a Russell style index based on its self-reported style where available, and otherwise based on its scaled rank on size and conditional value composite indicator.

portfolios.[9] A more comprehensive measure of value/growth orientation lowers the tracking volatility further to 8.7%.[10]

The active portfolios are concentrated stock groupings with a changing makeup and their returns contain a relatively large idiosyncratic component. They therefore provide tough challenges to track and the benchmarking procedures tend to be closely clustered in terms of their tracking error volatilities. Chan, Karceski, and Lakonishok (1999) provide additional perspective. They construct portfolios that are optimized under a tracking error variance criterion, and examine how the results change as they apply different models to forecast return covariance matrices. They find that models with varying degrees of complexity do not produce large differences in realized tracking error volatility out-of-sample. Placed in this context, the differences in tracking error volatility that we document across methods are material.[11]

To help identify the sources of the differences between methods' tracking error volatilities, Table 1 also reports how closely an active portfolio aligns with each benchmark in terms of several key features. For each stock in a portfolio, its rank on either size, book-to-market, or its composite value score is compared with the corresponding rank of its matching reference portfolio.[12] We calculate the simple mean of the absolute differences of these ranks across all stocks in the portfolio. For instance, when compared to equally weighted reference portfolios from independent sorts on size and book-to-market, the average managed fund has a mean absolute difference in size rank from its benchmark of 0.028.

---

[9] Many studies adopt the benchmarks developed by Daniel et al. (DGTW, 1997). They sort stocks into quintiles first by size, and then by book-to-market ratios within each size category; stocks are value-weighted in each of the 25 resulting portfolios. When we replicate our analysis with the DGTW reference portfolios, we find that, in general, the DGTW procedure does not dominate our version of sequential sorts using size and book-to-market. In particular, averaged across all the managed portfolios for the overall period, the DGTW yardsticks do about as well as the value-weighted benchmarks from independent sorts: tracking error volatility on average is 9.38% for DGTW compared to 9.35% for independent sorts. Results for the 1998–2000 subperiod, when the imbalance between large growth and large value becomes particularly problematic, bring out more clearly the advantage of the DGTW size-dependent sort. Tracking error volatility falls from 10.29% for value-weighted independently sorted benchmarks to 9.97% for DGTW. Nevertheless, the DGTW size quintile grouping is too coarse to capture well the behavior of large versus smaller firms: our size-conditional book-to-market method using finer size partitions generates lower tracking error volatility (8.97% on average for value-weighted benchmarks). Since the DGTW benchmarks do not do better than our procedures, we do not include them in the subsequent analysis.

[10] To assess the benefit from reduced tracking error volatility, consider the number of years required to declare an abnormal annual return of 4% to be reliably nonzero at the 10% significance level. This is roughly $(\frac{1.64\sigma}{4})^2$, where σ is the tracking error volatility. For σ of 10.37% from independently sorted control portfolios, for example, the required sample size is 18 years, compared to 13 years if the tracking volatility is 8.71%. In other words, the procedure with higher tracking volatility suffers an efficiency loss of 38% relative to the procedure with lower volatility.

[11] The F-statistic to test for equality of tracking error variances across the methods in Table 1 is 1.90 (p-value of 0.08).

[12] Ranks are calculated for all domestic common equities with coverage on the CRSP and Compustat databases. In July of each year, stocks are ordered and assigned ranks from zero (for the stock with the lowest value of the attribute) to one (for the stock with the highest value of the attribute). Similarly, a reference portfolio's attribute rank is the weighted average rank of its component stocks, with weights given by the beginning-of-period portfolio proportions.

Contrasting value-weighted and equally weighted versions of the benchmarks indicates that the former have an edge in matching the active portfolios' attribute ranks. The reduced differences help account for the lower tracking error volatilities produced by value-weighted reference portfolios.

The methods all perform comparably in terms of how closely they match the size and book-to-market features of the managed portfolios. However, they deviate more with respect to the managed portfolios' composite value indicator. The magnitude of the differences in the composite attribute ranks tends to line up with tracking error volatilities. Equally weighted benchmarks from independent sorts generate absolute differences on average of 0.089 and tracking error volatility on average of 10.37%. For equally weighted benchmarks matched on size and the composite indicator, the mean absolute difference is 0.017 and the tracking error volatility is 8.72%. The implication is that the procedure of matching portfolios only on size and book-to-market characteristics, which is customary in many academic studies, may overlook important sources of predictable variation in returns.

Nevertheless, even the size and value composite approach does not do much better than the Russell indexes based on tracking error. The latter method gives a tracking error volatility of 8.94% on average, despite the large mean absolute differences with respect to the portfolio characteristics. The Russell indexes are value-weighted portfolios with low volatility; book-to-market ratios are supplemented by long-term growth rate forecasts to assign stocks within a size category to value and growth subsets. These features of the Russell benchmarks may partly account for their relatively strong showing. Additionally, since the Russell indexes are widely used for evaluation purposes, many managers constrain themselves from being too out of line with respect to these benchmarks. For example, they may try to limit how far their portfolio weights deviate from the index weights, and they may try to stay close to the industry composition of the index. Note also that the indexes are based on relatively coarse breakdowns by size and value/growth orientation: for example, stocks within a size category (such as the largest 1000 stocks) are partitioned into only two groups (value and growth) so that they have roughly the same total market capitalization. As a result, the deviations with respect to characteristics can be sizeable.

Managers' track records diverge markedly during the 1998:Q1–2000:Q1 subperiod (panel B of Table 1).[13] As an illustration of how differences in the return behavior of equity asset classes are amplified during this period, in the case of independently sorted reference portfolios, the equally weighted benchmarks yield mean abnormal returns of 3.37%. Value-weighted versions of the same benchmarks generate mean abnormal returns of 0.65%. Comparisons

[13] The cross-sectional standard deviation of abnormal returns during this subperiod ranges from about 11% to 14% across the methods. By comparison, the standard deviations for the overall period are only about 4.5%.

Table 2
Performance (in percentage per year) of managed portfolios using alternative characteristics-based benchmarks, classified by investment style

(A) Large-capitalization portfolios

Method	Large-growth portfolios				Large-value portfolios				Large-cap portfolios	
	Abnormal return		Tracking error volatility		Abnormal return		Tracking error volatility		Mean abnormal return	Mean tracking error volatility
	Mean	Median	Mean	Median	Mean	Median	Mean	Median		
Panel A1: Full period, 1989:Q1–2001:Q4										
Annual independent size, BM										
Equal weight	4.03*	3.84	9.64	8.40	0.60	0.43	7.43	6.48	2.60	8.72
Value weight	2.11*	2.00	8.55	6.94	0.02	0.94	7.19	6.88	1.24*	7.98
Annual size, within-size BM										
Equal weight	2.97*	2.72	8.58	6.84	−0.25	−0.46	6.65	5.72	1.63*	7.78
Value weight	2.32*	2.25	7.99	6.22	−0.35	−0.82	6.64	5.54	1.21*	7.42
Annual size, value composite										
Equal weight	0.24	−0.23	7.78	7.00	0.96	0.91	5.64	4.84	0.54	6.89
Value weight	0.28	0.33	7.91	6.56	0.56	0.48	5.62	4.80	0.40	6.96
Quarterly size, within-size BM										
Equal weight	1.60*	1.37	8.10	6.68	−0.60	−0.73	6.32	5.40	0.69*	7.36
Value weight	1.47*	1.15	7.97	6.48	−0.64	−0.56	6.20	5.28	0.59*	7.24
Russell	3.28*	3.40	8.83	7.74	1.09	0.96	6.15	4.96	2.37*	7.72
Panel A2: 1998:Q1–2000:Q1										
Annual independent size, BM										
Equal weight	10.85	8.54	11.67	9.48	−3.35	−3.93	7.72	5.86	4.11	9.79
Value weight	7.09	3.99	10.76	7.78	−6.82	−7.09	8.01	6.90	0.49	9.45
Annual size, within-size BM										
Equal weight	10.69	7.01	10.88	8.60	−4.07	−4.58	6.79	5.64	3.68	8.94
Value weight	6.41	3.02	9.77	7.52	−5.51	−6.34	7.58	6.10	0.75	8.73
Annual size, value composite										
Equal weight	2.67	2.02	9.56	7.90	0.15	0.68	6.47	4.68	1.47	8.09
Value weight	1.52	1.86	9.64	7.08	−0.73	0.30	6.58	4.64	0.45	8.19
Quarterly size, within-size BM										
Equal weight	4.97	3.07	10.24	9.26	−4.86	−4.92	6.86	5.30	0.31	8.64
Value weight	4.78	2.30	10.01	8.46	−5.10	−4.83	6.83	5.28	0.09	8.50
Russell	3.12	0.70	9.93	8.40	−2.24	−1.69	6.48	5.44	0.58	8.30

(B) Small-capitalization portfolios

Panel B1: Full period, 1989:Q1–2001:Q4										
Annual independent size, BM										
Equal weight	2.89	1.04	15.52	14.86	−0.90	−1.14	11.80	11.00	1.00	13.66
Value weight	3.08	0.95	14.21	13.84	0.23	0.51	10.68	9.36	1.66	12.44
Annual size, within-size BM										
Equal weight	2.43	0.41	15.12	14.80	−1.52	−2.32	11.58	10.86	0.45	13.35
Value weight	2.69*	0.84	13.93	13.74	−0.53	−1.09	10.49	9.46	1.08	12.21
Annual size, value composite										
Equal weight	−1.91	−3.47	14.32	13.98	2.36	2.41	9.93	7.82	0.23	12.13
Value weight	−1.48	−2.86	14.23	13.60	2.76*	2.44	9.70	7.25	0.64	11.97
Quarterly size, within-size BM										
Equal weight	3.43*	1.67	13.95	14.04	−2.11	−2.76	10.66	9.80	0.66	12.30
Value weight	3.30*	1.47	13.72	13.08	−1.65	−2.52	10.19	9.04	0.82	11.96
Russell	7.28*	6.52	12.85	12.04	2.38*	2.13	7.73	6.84	4.83*	10.29
Panel B2: 1998:Q1–2000:Q1										
Annual independent size, BM										
Equal weight	13.94	12.90	17.36	15.56	−8.35	−9.51	9.42	8.86	0.77	12.67
Value weight	13.00	11.88	16.50	14.22	−7.78	−9.08	9.73	7.62	0.72	12.50
Annual size, within-size BM										
Equal weight	13.03	10.97	16.83	15.78	−8.11	−10.92	8.90	8.36	0.53	12.14
Value weight	11.95	8.43	15.71	15.92	−8.42	−11.45	9.66	8.10	−0.09	12.14
Annual size, value composite										
Equal weight	−3.08	−4.81	13.07	11.76	−0.76	−2.19	8.68	7.00	−1.71	10.48
Value weight	−1.44	−3.23	13.02	12.00	0.24	−1.12	8.78	6.90	−0.45	10.51
Quarterly size, within-size BM										
Equal weight	12.46	10.28	16.20	15.74	−9.42	−12.08	9.20	8.58	−0.47	12.06
Value weight	13.29	10.80	15.81	14.40	−8.04	−11.19	8.96	8.22	0.69	11.76
Russell	12.18	11.97	14.58	13.34	2.87	0.82	7.81	6.18	6.68	10.58

At the beginning of a quarter, every stock held in a managed portfolio is matched with a control portfolio based on its characteristics, using one of several procedures. The benchmark return is the weighted average of the quarterly buy-and-hold returns on the control portfolios. The procedure is repeated every quarter. A managed portfolio's performance is measured as its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4), and during 1998:Q1–2000:Q1. A portfolio's mean abnormal return is its annualized geometric mean return minus the annualized geometric mean return on the benchmark. A portfolio's tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the portfolio's return and the benchmark's return. For each performance measure, the arithmetic mean and median are provided over the cross-section of 199 managed portfolios in the sample period. For results based on the full sample period, an asterisk associated with the arithmetic mean abnormal return denotes that the time series mean of the quarterly equally weighted average return in excess of the benchmark across all available managed portfolios is at least two standard errors away from zero. Results are provided for large-capitalization (growth or value) portfolios in part A, and for small-capitalization (growth or value) portfolios in part B. A portfolio's investment style is based on its self-reported style where available, and otherwise based on its scaled rank on size and composite value indicator. The procedures to construct control portfolios are as follows. Under the independent sorting procedure, there are 25 control portfolios from the intersection of independent sorts by size (market value of equity) and BM (the ratio of book value of common equity to the market value of common equity). Under the size, within-size BM sort procedure, there are 28 control portfolios from sorts first by size, and then within each size category, by BM. In the size, value composite approach, a stock is given an overall ranking, conditional on its size group, based on book-to-market, dividend yield, cash flow yield, average earnings yield (based on the past year's net income, forecasted next year earnings, and forecasted two-year ahead earnings), and sales-to-price ratio. In these methods, the component stocks in a control portfolio are refreshed once a year at the end of June. In the quarterly size, within-size BM approach, the component stocks are refreshed at the beginning of each quarter. The return on a control portfolio is either the equally weighted or value-weighted average of the buy-and-hold returns on the component stocks. Each managed portfolio is also paired with a Russell style index corresponding to its investment style.

of tracking error volatilities in the later subperiod do not materially change the earlier conclusions from panel A.[14]

In summary, reference portfolios generated from different versions of the same methodology based on matching size and value/growth attributes deliver quite different verdicts about the performance of money managers. Benchmarks derived from independent sorts by size and book-to-market fare particularly poorly in terms of tracking active portfolio returns. A method that uses a finer partitioning of stocks into size brackets, and a more comprehensive measure of value/growth orientation, delivers lower tracking errors.

2.3 Results by investment style

Table 2 disaggregates the sample into subsets of managers who follow the same style. The style classification follows the same procedure as in Section 2.1.5. For the sake of brevity, we only present results for the four key styles: large value and growth, and small value and growth portfolios. Results are provided as well for the aggregated set of large value and growth portfolios (denoted large-cap) and aggregated small-cap portfolios.

The results from Table 2 generally buttress the overall conclusions from Table 1. Even when we narrow attention to managers who follow the same style, there are striking differences across methods in mean abnormal returns. To take the category of large growth portfolios as an example (part A), average levels of abnormal return run the gamut from a paltry 0.24% relative to equally weighted reference portfolios based on size and the composite value indicator, to a dazzling 4.03% based on equally weighted benchmarks from independent sorts on size and book-to-market. For the combined sample of large-growth and large-value portfolios, the range in abnormal returns across methods is 2.20%, while for the combined sample of small-stock portfolios in part B the range climbs to 4.60%.

Pronounced differences in mean abnormal return estimates come to the fore during the overheated 1998–2000 market. Benchmarks from independent sorts tend to produce large abnormal returns. Under this method, for example, large-growth managers outperform by 10.85% relative to equally weighted portfolios and 7.09% relative to value-weighted benchmarks. The true level of managerial skill in the sample is unknown, but average abnormal returns of this magnitude strain credulity. Reference portfolios based on size and the composite value indicator yield lower mean abnormal returns. In the case of large-growth managers, mean abnormal returns from this method are 2.67% (1.52%) for equally weighted (value-weighted) benchmarks.

---

[14] Comparing tracking error volatilities between the overall period (panel A) and the 1998–2000:Q1 subperiod (panel B) indicates little sign of change. This is misleading, however, because the overall period extends from the early 1990s when the investment industry generally was less concerned about tracking indexes. There is a decline in the overall level of tracking error volatility until the late 1990s when it jumps up. In the 1995–1997 period immediately preceding the subperiod in panel B, for example, tracking error volatilities for the sample average 6.58% across methods.

On the other hand, large-value managers severely underperform reference portfolios from independent sorts. Their mean abnormal return averages −3.35% under equally weighted benchmarks and −6.82% under value-weighted benchmarks. Abnormal returns of this magnitude are unlikely to reflect true performance, but rather cast doubt on the validity of treating size and book-to-market independently. In particular, the procedure tends to pair large-value managers with large-growth benchmarks. Since this reference group's return is a high hurdle to overcome during the late 1990s, large-value managers fare badly when compared to such an unrepresentative benchmark. (The following section elaborates on the extent of the resulting mismatch.) Adopting the size-conditional, composite value indicator paints a more accurate picture of a portfolio's value/growth style, yielding estimates of abnormal returns that are much less extreme. Mean abnormal returns are 0.15% under equally weighted benchmarks and −0.73% under value-weighted benchmarks.

Since the idiosyncratic component of returns is generally smaller for large stocks, cross-method comparisons of tracking error volatilities are likely to be more informative when applied to the large stock portfolios. Further, the bulk of institutional assets is concentrated in large-capitalization stocks. Accordingly, our discussion of the tracking error results in Table 2 focuses on the large growth and value managers.[15] Control portfolios based on independent sorts and book-to-market ratios are generally associated with the highest tracking error volatilities. Within-size sorts by a more comprehensive measure to profile value/growth reduce mean tracking error volatility. In the combined large-stock manager sample, for example, the improvement in tracking volatility is from 8.72% for equally weighted independently sorted benchmarks to 6.89% for equally weighted benchmarks from sorts by size and the composite value indicator. The corresponding reduction for value-weighted benchmarks is from 7.98% to 6.96%.

In the case of small-stock portfolios (part B), idiosyncratic return volatility is higher and smudges the differences across methods in tracking error volatilities. Nevertheless, the overall conclusions from part A are unaffected. For example, it is still the case that the independent sort procedure performs poorly with respect to tracking ability compared to the size, composite value approach.

For both large- and small-stock portfolios, the tracking error volatilities convey the message that procedures based on book-to-market as the sole measure of value/growth orientation perform poorly. Evidently, book-to-market misses important information about return comovement. Treating as identical two similarly-sized firms that have the same book value turns a blind eye to differences along other important dimensions, such as profitability, for instance.

[15] The time clustering-adjusted F-statistic to test whether the equally weighted portfolio of large-capitalization managers has equal tracking error volatilities across methods is 2.23 with a p-value of 0.04.

3. Interpreting the Evidence from Characteristic-Matching Methods

In this section, we trace the sources of the differences in results across benchmarking techniques, with the objective of identifying the relative merits of each method. We do this in several ways. First, we apply the benchmarking procedures to a set of passive portfolios. This lets us see how the methods fare in a controlled setting where there is no managerial skill, the composition of the portfolios is fairly stable, and the idiosyncratic return component is small. These conditions help to bring the benchmarking methods' performance into sharper focus. Second, we provide further details on the characteristics of the baseline portfolios from different methods.

3.1 Results for passive indexes

Table 3 provides results when we take as our pseudo-active portfolios eight Russell style indexes: the Russell top 200 growth and value indexes; the Russell mid-cap growth and value indexes; the Russell 1000 growth and value indexes; and the Russell 2000 growth and value indexes. These indexes are commonly used in the investment industry to evaluate managers.[16] Table 3 also reports on the simple average over the eight indexes of the abnormal return, the absolute abnormal return so that the positive and negative excess returns do not cancel out, and the tracking error volatility.

The Russell indexes represent large, well-diversified portfolios, which are, when compared to the managers in our sample, less concentrated with a more stable composition. Accordingly, abnormal returns on the indexes should not differ markedly from zero and the benchmarks should track the indexes closely. This potentially affords more room for the different methods to stand out clearly from one another. Even with these relatively well-behaved passive portfolios and long sample periods, however, the methods can yield quite different conclusions about abnormal returns. In the case of the Russell 1000 growth index, for example, the methods report net-of-benchmark returns that range from a low of -1.66% to a high of 1.08%.

Taking the benchmarking methods to unmanaged indexes that are well diversified with relatively fixed make-up succeeds in spreading out tracking error volatility across methods. In particular, the independent sort procedure stands out for its poor covariation with the broad-based passive Russell indexes: in seven out of the eight series this method yields the largest tracking error volatilities across methods. On the other hand, sorts by size and then by the composite value measure yield benchmarks that covary strongly with the indexes.

[16] Each index refers to growth or value stocks within a given size category. The largest 200 stocks by market capitalization constitute the top 200, while the next 800 make up the mid-capitalization group. The Russell 1000 comprises these two groups. The Russell 2000 comprises the next largest 2000 stocks. Within each size category, stocks are ranked by a score based on book-to-market ratio and analysts' estimates of long-term earnings growth rates. Stocks are then assigned to value or growth partitions such that half of the total market capitalization of the size category is in each partition. The return on the index is the value-weighted average of the component stocks' returns, where the weights are adjusted for cross-ownership and privately held shares.

Table 3
Performance (in percentage per year) of Russell style indexes using alternative characteristics-based benchmarks

		Annual independent size, BM		Annual size, within-size, BM		Annual size, value composite		Quarterly size, within-size, BM	
		Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight
				Panel A: Full period, 1989:Q1–2001:Q4					
Top 200 Growth	Abnormal return	1.05	−0.18	0.48	0.33	−1.60	−1.12	0.20	0.46
	Tracking error	5.68	3.68	4.06	3.00	3.42	2.12	2.58	3.34
Top 200 Value	Abnormal return	1.70	1.52	1.14	0.97	0.29	−0.17	0.69	0.61
	Tracking error	6.74	4.64	3.82	3.76	2.72	2.58	3.04	3.12
Mid-cap Growth	Abnormal return	1.23	0.88	1.91	1.19	−1.79	−2.04	2.33	2.78*
	Tracking error	8.38	9.62	6.72	5.72	4.28	4.36	6.44	6.16
Mid-cap Value	Abnormal return	0.35	0.76	0.16	0.63	0.28	0.31	0.70	0.55
	Tracking error	3.16	3.86	3.14	2.72	2.84	2.56	2.38	2.68
R1000 Growth	Abnormal return	1.08	0.07	0.85	0.59	−1.66	−1.28	0.70	0.99
	Tracking error	4.92	4.08	3.88	3.10	3.06	2.20	2.70	2.98
R1000 Value	Abnormal return	1.29	1.31	0.85	0.87	0.31	−0.02	0.72	0.64
	Tracking error	5.26	4.10	3.40	3.32	2.26	2.16	2.66	2.72
R2000 Growth	Abnormal return	−0.04	0.48	0.23	0.31	−3.59*	−3.69*	−0.40	−0.15
	Tracking error	7.08	6.02	6.80	5.58	4.60	4.24	5.38	5.86
R2000 Value	Abnormal return	−0.08	1.11	0.29	0.54	1.02	1.19	−0.02	−0.06
	Tracking error	3.78	3.54	3.70	3.66	4.06	3.86	3.42	3.36
Average	Abnormal return	0.82	0.74	0.74	0.68	−0.84	−0.85	0.62	0.73
	Absolute abnormal return	0.85	0.79	0.74	0.68	1.32	1.23	0.62	0.73
	Tracking error volatility	5.64	4.94	4.44	3.86	3.41	3.01	3.58	3.78
				Panel B: 1998:Q1–2000:Q1					
Top 200 Growth	Abnormal return	7.56	3.72*	6.52	3.50	1.78	1.25	1.23	1.19
	Tracking error	5.78	2.42	5.22	2.84	3.46	2.14	1.80	2.70
Top 200 Value	Abnormal return	0.74	−4.15	−2.46	−4.96	1.96	0.78	−3.36	−2.04
	Tracking error	5.42	5.52	2.66	4.56	2.00	2.64	3.56	3.42
Mid-cap Growth	Abnormal return	16.36	13.31	15.39*	8.36	4.00	2.95	14.01*	14.00*
	Tracking error	14.24	13.94	11.78	7.96	6.76	6.98	10.78	10.84

(*Continued overleaf*)

Table 3
(*Continued*)

		Annual independent size, BM		Annual size, within-size, BM		Annual size, value composite		Quarterly size, within-size, BM	
		Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight	Equal weight	Value weight
Mid-cap Value	Abnormal return	−5.96*	−8.44*	−4.11*	−4.77*	0.34	0.19	−3.70*	−4.21*
	Tracking error	2.04	3.26	2.08	3.06	3.36	3.20	2.14	2.08
R1000 Growth	Abnormal return	8.90*	5.31*	8.09*	4.33	2.19	1.54	3.52*	3.46*
	Tracking error	5.32	3.74	5.54	3.32	2.92	2.52	2.60	2.40
R1000 Value	Abnormal return	−1.57	−5.64	−3.02*	−4.91	1.33	0.48	−3.50	−2.79
	Tracking error	4.22	4.68	2.20	4.04	1.76	1.98	2.92	2.80
R2000 Growth	Abnormal return	8.44	6.94	6.58	5.65	−4.54	−3.75	5.62	5.16
	Tracking error	10.98	9.00	11.20	7.76	6.02	5.42	8.62	10.82
R2000 Value	Abnormal return	−5.49*	−5.70*	−6.67*	−7.81*	−0.76	−0.55	−7.39*	−7.81*
	Tracking error	3.10	4.06	2.92	4.18	5.14	5.24	3.10	2.48
Average	Abnormal return	3.62	0.67	2.54	−0.08	0.79	0.36	0.80	0.87
	Absolute abnormal return	6.88	6.65	6.61	5.54	2.11	1.44	5.29	5.08
	Tracking error volatility	6.39	5.83	5.45	4.72	3.93	3.77	4.44	4.69

At the beginning of a quarter, every stock in a Russell style index is matched with a control portfolio based on its characteristics, using one of several procedures. The benchmark return is the weighted average of the quarterly buy-and-hold returns on the control portfolios. The procedure is repeated every quarter. For each index, statistics are provided for its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4), and during 1998:Q1–2000:Q1. The mean abnormal return is the annualized geometric mean return on the index minus the annualized geometric mean return on the benchmark. Tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the index return and the benchmark's return. An asterisk denotes that the mean abnormal return is more than two time-series standard errors away from zero. The indexes are: the Russell top 200 (value and growth); the Russell mid-cap (value and growth); the Russell 1000 (value and growth); and the Russell 2000 (value and growth). The performance measures under each procedure are also averaged across the eight indexes and reported at the bottom of each panel. The procedures to construct control portfolios are as follows. Under the independent size, BM procedure, there are 25 control portfolios from the intersection of independent sorts by size (market value of equity) and BM (the ratio of book value of common equity to the market value of common equity). Under the size, within-size BM sort procedure, there are 28 control portfolios from sorts first by size, and then within each size category, by BM. In the size, value composite approach, a stock is given an overall ranking, conditional on its size group, based on book-to-market, dividend yield, cash flow yield, average earnings yield (based on the past year's net income, forecasted next year earnings, and forecasted two-year ahead earnings), and sales-to-price ratio. In these methods, the component stocks in a control portfolio are refreshed once a year at the end of June. In the quarterly size, within-size BM approach, the component stocks are refreshed at the beginning of each quarter. The return on a control portfolio is either the equally weighted or value-weighted average of the buy-and-hold quarterly returns on the component stocks.

Averaged across all the indexes, this method produces tracking error volatilities of 3.41% for equally weighted reference portfolios, compared to 5.64% for baseline portfolios from independent sorts.[17] Restated in terms of the number of years necessary to declare a hypothetical mean abnormal return of 4% to be statistically significant at the 10% level, the independent sort procedure would require 5.35 years while the size, composite value approach requires only 1.95 years.

The eye-catching differences across the methods during the 1998:Q1–2000:Q1 subperiod (panel B of Table 3) highlight the shortcoming of book-to-market as a summary measure of value/growth style. Baseline portfolios that measure value/growth orientation solely by book-to-market frequently give rise to implausibly large abnormal returns. For example, the abnormal return is 5.31% for the Russell 1000 growth index, and −5.64% for the Russell 1000 value index, under the value-weighted, independent sort procedure. Abnormal returns are generally closer to zero when judged against reference portfolios that take other criteria into consideration when classifying stocks as value or growth. With the capitalization-weighted size, value composite method, the abnormal return is 1.54% for the Russell 1000 growth (0.48% for the Russell 1000 value) index.

3.2 Features of characteristic-matched portfolios

We concentrate on the features of reference portfolios from independent sorts on size and book-to-market. This set of benchmarks is extensively used in the research literature, in no small part because the data are easily accessible from Ken French's website.

Table 4 reports on the percentage of market capitalization accounted for by each of the 25 control portfolios from independent sorts. The distribution is calculated at the beginning of each quarter from the first quarter of 1989 to the last quarter of 2001. The results are averaged over quarters, and are provided for four subperiods: 1989:Q1–1994:Q4, 1995:Q1–1997:Q4, 1998:Q1–2000:Q1, and 2000:Q2–2001:Q4. To conserve space, we present results for only the 1989:Q1–1994:Q4 and 1998:Q1–2000:Q1 subperiods.

Not surprisingly, the top quintile of stocks accounts for the bulk of market capitalization. The discomfiting feature of the independent sort procedure, however, is the highly uneven split between growth and value stocks within the large capitalization subset. In the first subperiod (panel A1 of Table 4), the large-growth category represents 25.57% of the total value of listed domestic U.S. stocks while the large-value group makes up only 4.76%. As a result of the steep run-up in the prices of large-growth firms during the market boom, the relative importance of this group climbs in the late 1990s. Large-growth stocks' weight

[17] As another comparison, the tracking error volatility for the Russell indexes when benchmarked against the Daniel et al. (1997) size- and book-to-market matched portfolios is on average 4.60%. This is close to the result from value-weighted independently sorted benchmarks (4.94%). On the other hand, our value-weighted, sequentially sorted size and book-to-market portfolios yield lower tracking error volatility (3.86% on average).

Table 4
Comparison of distribution of market capitalization across size and book-to-market control portfolios

Panel A: Based on independent size, book-to-market breakpoints

Panel A1: 1989:Q1–1994:Q4

	1 (growth)	2	3	4	5 (value)
1 (large)	25.57	16.25	14.35	10.03	4.76
2	4.25	3.17	2.96	2.74	1.64
3	2.21	1.52	1.43	1.20	0.91
4	1.29	0.94	0.90	0.69	0.58
5 (small)	0.77	0.46	0.38	0.41	0.61

Panel A2: 1998:Q1–2000:Q1

	1 (growth)	2	3	4	5 (value)
1 (large)	46.36	18.07	8.18	4.38	3.39
2	3.27	2.26	1.75	1.36	1.12
3	1.55	1.01	0.90	0.71	0.45
4	0.87	0.63	0.60	0.53	0.34
5 (small)	0.52	0.39	0.43	0.48	0.47

Panel B: Based on size, within-size book-to-market breakpoints

Panel B1: 1989:Q1–1994:Q4

	1 (growth)	2	3	4	5 (value)
Top 75 (large)	15.97		14.70		14.08
Next 125	3.97	3.94	3.98	3.90	3.71
Next 300	3.59	3.62	3.72	3.73	3.58
Next 500	1.99	1.96	1.95	2.01	2.01
Next 1000	1.11	1.14	1.15	1.13	1.03
Rest (small)	0.41	0.42	0.41	0.41	0.37

Panel B2: 1998:Q1–2000:Q1

	1 (growth)	2	3	4	5 (value)
Top 75 (large)	20.66		16.15		12.76
Next 125	3.61	3.63	3.59	3.67	3.08
Next 300	3.01	3.05	3.03	2.93	2.99
Next 500	1.66	1.73	1.72	1.73	1.68
Next 1000	1.20	1.20	1.18	1.18	1.10
Rest (small)	0.73	0.75	0.73	0.69	0.59

At the end of June each year from 1989 to 2001, the market value of common equity (as of June-end) and the ratio of book value of common equity (from the prior fiscal year) to market value of common equity (from December of the prior calendar year) is computed for each domestic U.S. common stock listed on the NYSE/AMEX/NASDAQ markets. Based on these values relative to breakpoints, each stock is placed in a category of size and book-to-market. The total market value in each category relative to aggregate market value is calculated each year and averaged over periods. Breakpoints are calculated in two ways. In the first way (panel A) the breakpoints for size are quintile values determined from sorting NYSE stocks only; the breakpoints for book-to-market are NYSE quintile values obtained from an independent sort of all domestic common stocks each year. The total number of categories is 25. In the second way (panel B), there are six categories of size: top 75 by market capitalization, the next 125, the next largest 300, next 500, next 1000, and the remainder ranked by market value of equity. Within the largest 75 stocks, firms are ranked by book-to-market ratio and placed in one of three groups with equal numbers of firms each. Within each of the other five groups by size, firms are ranked by book-to-market and placed in one of five groups with equal number of firms in each. The total number of categories is 28.

averages 46.36% in the 1998:Q1–2000:Q1 subperiod; conversely, large-value stocks shrink in importance to only 3.39% of capitalization.[18]

To rephrase the argument, the percentage amount in the cells of Table 4 can be interpreted as the distribution of assets across investors of different styles. From this perspective, the independent sort procedure suggests that in the late 1990s, large-capitalization growth investors command as much as 14 times the assets of large-capitalization value managers. In fact, the distribution of clients' mandates is typically more evenly divided between value and growth. Simply put, investors' behavior does not conform to the classification produced by independent sorts on size and book-to-market.

Panel B of Table 4 provides the corresponding distribution of market capitalization for the classification based on size-conditional book-to-market breakpoints. In comparison to panel A, the split of large stocks into growth and value partitions is more balanced. The large-growth category is much less dominant, and its relative importance is more stable across subperiods. Large-growth stocks contribute 20.66% of market capitalization in the 1998:Q1–2000:Q1 subperiod, for example, compared to 15.97% in the first subperiod.

Given the lopsided distribution produced by independent size and book-to-market breakpoints, the resulting benchmarks are heterogeneous portfolios that may be poorly aligned with more focused, active portfolios. Table 5 documents the extent of the problem. Following up on the comparisons of the previous table, we single out the large-growth benchmark portfolios from either independent sorts, or from the size-conditional book-to-market classification. Various attributes of each portfolio are reported in Table 5 to assess where they fall along the value/growth spectrum. To ease comparison, we express each attribute as equidistant percentile ranks from zero to one, so a stock with the highest value of the attribute (the most value-oriented stock) receives a rank value of one while the stock with the lowest value of the attribute (the most growth-oriented stock) receives a rank value of zero. Percentiles of the distribution of attribute ranks are calculated over stocks in the portfolio and are then averaged over all quarters, or over the 1998:Q1–2000:Q1 subperiod.

In many studies, a stock is considered as value or growth based on its book-to-market ratio, so this is the first characteristic we consider. As other indicators of a stock's value/growth profile, we also consider: cash flow yield, dividend yield, earnings yield, and sales-to-price ratio.

Every measure of value/growth orientation exhibits large variation within the large-growth benchmark from independent sorts (panel A of Table 5). The earnings yield ranks of stocks in this group extend from 0.1062 at the 10th percentile to 0.5242 at the 90th percentile. In comparison, the large-growth

[18] The composition of the categories is determined once a year (at the end of June). The average for 1998:Q1–2000:Q1 thus misses the peak of the market boom since the last classification occurs in mid-1999. Accordingly, some of the effects of the 1998–2000 price run-up are picked up only in later years' classifications. On average over the 2000:Q2–2001:Q4 subperiod, for instance, the weight of large-growth stocks is 61.52% while the weight of large-value stocks is 1.70%.

Table 5
Characteristic ranks of large-growth benchmark portfolios

	Independent size, book-to-market breakpoints					Comparison large-growth group				
Characteristic	10%	25%	Median	75%	90%	10%	25%	Median	75%	90%
Panel A: Full period, 1989:Q1–2001:Q4										
Book-to-market	0.0357	0.0733	0.1295	0.1926	0.2509	0.0330	0.0505	0.0843	0.1167	0.1449
Cash flow yield	0.0829	0.1494	0.2417	0.3479	0.4789	0.0796	0.1239	0.1842	0.2478	0.3105
Dividend yield	0.0550	0.1512	0.3299	0.4957	0.6291	0.0546	0.1564	0.3208	0.4154	0.4912
Earnings yield	0.1062	0.1908	0.2934	0.4026	0.5242	0.1081	0.1631	0.2434	0.3223	0.3832
Sales to price	0.0725	0.1229	0.2251	0.3664	0.5324	0.0720	0.1075	0.1510	0.2503	0.3187
Value rank	0.0556	0.1250	0.2564	0.3978	0.5393	0.0308	0.0709	0.1557	0.2547	0.3366
Panel B: 1998:Q1–2000:Q1										
Book-to-market	0.0211	0.0474	0.0921	0.1529	0.2087	0.0123	0.0201	0.0378	0.0632	0.0780
Cash flow yield	0.0607	0.1032	0.1905	0.3077	0.4760	0.0355	0.0576	0.0958	0.1451	0.2054
Dividend yield	0.0381	0.1205	0.2906	0.4856	0.6544	0.0287	0.0865	0.2356	0.3223	0.3839
Earnings yield	0.0695	0.1338	0.2254	0.3312	0.4540	0.0527	0.0812	0.1502	0.2131	0.2590
Sales to price	0.0550	0.1039	0.2011	0.3522	0.5161	0.0428	0.0602	0.0967	0.1533	0.2081
Value rank	0.0726	0.1558	0.3024	0.4769	0.6541	0.0301	0.0763	0.1661	0.2709	0.3449

Characteristics, expressed as percentile rank values from zero to one, are reported for two sets of stocks selected at the end of June each year from 1989 to 2001. The first set comprises stocks classified as large growth from independent sorts by market value of common equity and the ratio of book value to market value of common equity. The second set comprises large stocks (the largest 75 based on market capitalization) classified as most growth-oriented based on an overall indicator of value/growth orientation. The overall indicator is the average of a stock's percentile rank on each of five variables (book-to-market, cash flow yield, dividend yield, sales-to-price, and average earnings yield), where ranks are relative to firms in the same size category. A stock's characteristic rank is obtained at the beginning of each quarter by ranking all eligible U.S. listed domestic common stocks by the value of the characteristic and assigning its percentile rank such that the stock with the lowest (highest) value of the attribute has a rank of zero (one). For each set of stocks, percentiles of the distribution of characteristic ranks are calculated each quarter and averaged over the entire sample period (1989:Q1–2001:Q4) and for the 1998:Q1–2000:Q1 subperiod. The characteristics are: book value of common equity relative to market value of common equity; cash flow yield (operating income before depreciation relative to market value of firm, measured as total assets minus book value of equity and accounts payable, plus market value of common equity); dividend yield (cash dividends to common equity relative to market value of equity); earnings yield (net income available to common equity relative to market value of equity); sales-to-price (net sales relative to market value of equity); and a composite value indicator. The composite value indicator is the rescaled average of a stock's percentile rank, relative to stocks in the same size category, of book-to-market, cash flow yield, dividend yield, sales-to-price ratio, and average earnings yield (average of percentile ranks of prior year net income relative to market capitalization, consensus forecast of next year earnings relative to price, consensus forecast of two-year ahead earnings relative to price). All accounting variables are measured as of the prior fiscal year, while market value of equity is measured in December of the prior calendar year.

benchmark based on within-size breakpoints for book-to-market comprises a more homogeneous collection of stocks. Their corresponding earnings yield ranks run from 0.1081 to 0.3832.[19]

Further, the large-growth reference portfolio from independent classifications embraces many stocks that would not generally be considered very growth-oriented. Based on the overall value indicator, for instance, the 75th percentile of the distribution is 0.3978. Therefore, a quarter of the stocks in the portfolio score above the fourth decile in terms of value/growth tilt within their size partition. In short, the large-growth benchmark from an independent

[19] Note that the independent sort procedure uses New York Stock Exchange breakpoints for size and book-to-market. However, our percentile ranks on book-to-market are determined relative to the cross-section of all listed domestic common stocks. As a result, stocks classified as large growth under independent sorts do not necessarily have ranks that fall below 0.2.

size, book-to-market classification does not faithfully mirror the equity class it purports to depict. Stated differently, many of the stocks that a large-value manager would hold in practice are classified as large-growth stocks under an independent sort procedure. The result of this scheme is to pair off a large-capitalization value-oriented active manager with an unrepresentative reference portfolio.

The heterogeneity is exacerbated during the late 1990s (panel B of Table 5). Within the large-growth benchmark from independent classifications, the spread between the 90th and 10th percentiles of the distribution of the composite value indicator is 0.5815. A quarter of the stocks in the portfolio have a value indicator rank in excess of 0.4769. On the other hand, the size-conditional book-to-market classification produces a benchmark portfolio that is more tightly focused in terms of its large-growth orientation. The difference between the 90th and 10th percentiles of the value composite score within this group is only 0.3058.

4. Regression-based Benchmarks

Matching each stock in a managed portfolio against a control portfolio has the advantage of yielding potentially more accurate measures of expected future returns. The disadvantage is that the data requirements are more burdensome, since the portfolio manager's holdings at the beginning of the period must be known. The alternative is to work with the realized returns on the managed portfolio.

4.1 Three-factor time-series regressions

Fama and French (1993) develop a three-factor model of the form

$$r_{pt} - r_{ft} = \alpha_p + \beta_p(r_{mt} - r_{ft}) + h_p HML_t + s_p SMB_t + \epsilon_{pt}, \qquad (1)$$

where $r_{pt} - r_{ft}$ is the return on portfolio p in period t in excess of the risk-free rate and $r_{mt} - r_{ft}$ is the excess return on the market. HML_t is the return on a zero investment factor-mimicking portfolio that is long on value stocks and short on growth stocks; similarly, SMB_t is the return on a zero investment factor-mimicking portfolio that is long on small stocks and short on large stocks. In the absence of stock selection ability, α_p should equal zero.

The Fama and French (1993) size and value/growth factors are measured as the differences between the returns of extreme portfolios from independent sorts on size and book-to-market. To follow up on the evidence in the prior sections suggesting that independent sorts tend to yield heterogeneous stock clusters, we develop alternative mimicking portfolios. In particular, size is the difference between the value-weighted return on large stocks (the 200 largest companies by equity market capitalization) and the value-weighted return on small stocks (the 1000 stocks ranked below 1000 when ordered by size). Similarly, because

book-to-market equity may be an incomplete description of a stock's value/growth profile, we use our composite value measure to define the value/growth factor. First, we calculate within each size cohort the difference in quarterly value-weighted returns between the top and bottom third of stocks ranked by the composite. The spread is then averaged across size classes to yield the time series of value factor returns.

Equation (1) accounts for the effects of size and value/growth separately, so the average benchmark return on a portfolio adds a reward for smallness and a reward for value (in addition to the compensation for market exposure). This may adequately describe return behavior over long periods, but it may not be an innocuous assumption over the short horizons where performance is typically measured. Consider, for example, a portfolio manager who concentrates in small-value stocks, that is, who loads heavily on smallness and on value. This investor will be held to a high predicted return when small stocks outperform large stocks. However, the model posts a high expected return for smallness even if the only reason small stocks do well is because small-growth stocks outperform. In this circumstance, the hurdle is set too high for portfolios of small-value stocks and too low for small-growth stock portfolios. Such an event occurs, for example, in the first quarter of 2000, when small stocks (as measured by the Russell 2000 index) earned a return of 7.08%. This exceeds the return in the same quarter of 4.37% on the Russell 1000 index of large stocks. In the small-stock cohort, however, small-growth stocks in the Russell 2000 growth index posted a larger return (9.29%) than small-value stocks in the Russell 2000 value index (3.82%).

4.2 Effective asset mix regressions

The three-factor regression model appears extensively in academic research. In the investment industry, an alternative regression-based benchmarking approach, due to Sharpe (1992), is more popular. An active manager is viewed as choosing stocks from equity subsets that vary across the size spectrum, and across the value/growth spectrum. The return on the manager's portfolio can thus be allocated into components corresponding to the return on each subset. Any differential return reflects the manager's skill.

We apply Sharpe's (1992) effective asset mix approach by estimating constrained regressions of the form

$$r_{pt} = \sum_{j=1}^{K} \gamma_{pj} I_{jt} + \upsilon_{pt}, \tag{2}$$

where I_{jt} are the returns at time t on the equity subclasses. The coefficients γ_{pj}, $j = 1, \ldots, K$, represent the proportions of portfolio p that are invested in each of the K classes. Since the equity managers in our sample are limited to long positions in stocks, we prevent estimating counterfactual coefficients by imposing the constraints that each $\gamma_{pj} \geq 0$, $j = 1, \ldots, K$, and $\sum_{j=1}^{K} \gamma_{pj} = 1$.

Part of the popularity of Sharpe's (1992) approach stems from its ease of interpretation, since the coefficients can be readily interpreted as portfolio weights. Importantly, Equation (2) uses the information in the returns to each distinct equity asset class. Consequently, it does not share the three-factor model's shortcoming when value and growth stocks behave differently across size cohorts.[20]

We use six equity style classes in Equation (2): large value and growth, mid-cap value and growth, as well as small value and growth. The returns on these classes are measured as the performance of either the Wilshire Target Indexes, the value-weighted benchmark portfolios from independent sorts that underly the Fama-French (1993) time-series factors, or the value-weighted reference portfolios from within-group sorts by size (small, mid-, and large-cap) and then the composite value measure (value and growth).[21]

4.3 Cross-sectional regression-based benchmarks

Empirical research on asset-pricing models fits regressions of returns on attributes such as beta, size, and book-to-market (see, for example, Chan, Hamao, and Lakonishok 1991 and Fama and French 1992). The thrust of this logic is that the fitted return from such a model can serve as the benchmark for an active portfolio, given the attributes of the stock held by the manager.

We formulate this argument as follows. Each quarter we estimate the following cross-sectional regression:

$$r_{it} = \lambda_{0t} + \sum_{j=1}^{L} \lambda_{jt} X_{jt} + \nu_{it}, \tag{3}$$

where r_{it} is the return of stock i over quarter t while X_{jt} are stock attributes at the beginning of the quarter. Given the estimates of the coefficients λ_{jt}, $j = 0, \ldots, L$ and the attributes of a stock, we calculate its fitted return from Equation (3). The benchmark return for an active portfolio is then the

---

[20] In the three-factor model, the mimicking portfolios for size and value are linear combinations of the returns on the equity subclasses, as is the market portfolio. Substituting these definitions into the factor model Equation (1) yields a regression of managed portfolio returns on all the underlying equity subclass returns, with restrictions on the coefficients. For example, the portfolio's coefficient on the return to the small-cap value subclass, SV, can be written as $\frac{1}{3}s_p + \frac{1}{2}h_p + \beta_p\omega_{SV}$, where s_p is managed portfolio p's loading on the size factor, h_p is its loading on the value factor, β_p is its market beta, and ω_{SV} is the capitalization weight of the small-cap value subclass relative to the market. Since the effective asset mix model corresponds to the full regression, it should produce lower tracking error volatilities, at least in-sample, if the nonnegativity and summation constraints on the weights in the Sharpe (1992) style regression are not inconsistent with the data. Note that another restriction of the three-factor model is that the sum of the coefficients over equity subsets equals the portfolio's market beta.

[21] Since we follow the standard practice of constraining the regression coefficients in Equation (2) to fall between zero and one, a manager cannot follow an estimated investment style that is more aggressive than any of the indexes. To give the effective asset mix procedure a fair chance to capture the entire span of manager styles, we select indexes that are relatively extreme along the value/growth spectrum. This argues against the Russell indexes, where there are only two value/growth categories, and these can overlap. Instead, the Target style indexes, produced by Wilshire Associates, are concentrated passive portfolios constituting stocks that clearly conform to high-growth or high-value features within a size bracket. Multiple distinct criteria are used to assign stocks to value and growth categories.

weighted average of the fitted returns of the stocks held by the manager using beginning-of-quarter investment weights.

Equation (3) is well known and extensively applied in financial research. In addition, it is the backbone of several performance evaluation and attribution systems that are widely used in the investment industry (see, for example, BARRA 1990). The model can be interpreted as a linear factor model for returns, where stock attributes are assumed to be accurate measures of exposures to the underlying factors. The measured coefficient λ_{jt} is an estimate of the realization of factor j in quarter t.

Daniel and Titman (1997) find that a simple linear or log-linear model does not fully capture the association between returns, size, and book-to-market. To give the cross-sectional regression a fair test in capturing the behavior of returns, we employ a specification that is parsimonious and reasonably robust. We include the key variables that have been found in the literature to be important determinants of the cross-section of average stock returns: size, book-to-market, cash flow to firm value, dividend yield, earnings yield, sales-to-price ratio, past six-month return, and industry dummy variables. The effect of firm size is captured through a set of five indicator variables. Depending on where a stock's market capitalization falls in the size distribution of NYSE firms, one of these indicators takes the value of one and the others are zero. The ranges are: the top 5% of size; from the 80th to the 95th percentile; between the 50th and 80th percentiles; between the 25th and 50th percentiles; and firm size below the 25th percentile. These cohorts are meant to partition stocks into subsets roughly corresponding to the equity investment domains of interest to managers. To mitigate problems with extreme values of the characteristics, we use percentile rank values of the accounting attributes (from zero for the lowest to one for the highest). Prior six-month returns for a stock are measured over a period ending one month before the return measurement month. The industry dummy variables are based on the Fama-French (1997) classification.

Equation (3) is typically applied in contexts where the objective is to uncover the determinants of returns over relatively long horizons. Using it to pin down the behavior of short-horizon returns such as a month or a quarter, as is done in practice, may be more treacherous. As a specific issue, the linear specification of the model assumes that the impact of a variable such as earnings yield is uniform across its entire range of values. This is a questionable assumption over short horizons.

4.4 Results

The results for regression-based benchmarks are reported in Table 6 for all managers. The return predicted for a portfolio in a given quarter is based on that quarter's realizations of the regressors along with the estimated loadings from either Equations (1), (2), or (3). Loadings are estimated over two sample periods: either the entire return history of an active portfolio, or its history excluding the quarter under evaluation. Fitting the regressions to the manager's

Table 6
Performance (in percentage per year) of managed portfolios using alternative regression-based benchmarks

		Fama-French three-factor model		Market, size, value composite factor model		Effective asset mix regressions with: Wilshire indexes		Effective asset mix regressions with: Independent sort size, BM portfolios		Effective asset mix regressions with: Size, value composite portfolios		Cross-sectional regression on attributes
		All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	
						Panel A: Full period, 1989:Q1–2001:Q4						
Abnormal return	Mean	2.19	2.64	3.64	3.67	3.03	3.09	1.49	1.48	2.02	2.26	−1.97
	Median	1.61	1.87	3.27	3.23	2.61	2.88	1.03	0.94	2.03	2.32	−1.78
	t-stat	5.22	4.80	6.44	5.47	3.28	3.10	3.02	2.75	2.45	2.06	−1.47
Tracking error vol	Mean	7.94	10.54	7.02	8.33	7.93	8.75	7.72	8.51	7.66	8.91	10.60
	Median	6.56	8.48	5.88	6.86	6.02	7.04	6.08	6.90	6.34	7.56	9.48
						Panel B: 1998:Q1–2000:Q1						
Abnormal return	Mean	2.94	4.29	4.45	4.45	8.08	8.37	3.12	3.17	4.19	4.67	−5.94
	Median	1.27	0.98	3.54	3.66	5.25	5.43	0.98	1.13	3.82	4.11	−5.26
Tracking error vol	Mean	9.63	14.77	7.54	9.08	8.78	9.70	8.45	9.52	8.66	10.07	13.19
	Median	7.86	11.72	6.20	7.34	6.32	7.46	6.78	7.68	7.20	8.30	11.22

Each quarter, a managed portfolio's benchmark return is the fitted value from one of a variety of regression models. A managed portfolio's performance is measured as its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4) and during 1998:Q1–2000:Q1. A portfolio's mean abnormal return is its annualized geometric mean return minus the annualized geometric mean of the fitted benchmark returns. A portfolio's tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the portfolio's return and the benchmark's return. For each performance measure, the arithmetic mean and median are provided over the cross-section of 199 managed portfolios in the sample period. The reported *t*-statistic is for the hypothesis that the time-series mean of the quarterly equally weighted average return in excess of the benchmark across all available managed portfolios equals zero. In panels A and B, the regression uses all quarters over the full period 1989–2001, or excludes the current quarter to estimate coefficients. The estimated slope coefficients, along with the realized values of the regressors in the current quarter, are used to generate the fitted value (any intercept term is suppressed). In the Fama-French (1993) three-factor model, the regressors are the market excess return and returns on mimicking portfolios for size and book-to-market, *SMB* and *HML*. In the market, size, and value composite factor model, the regressors are the market excess return, the difference between the value-weighted return on the largest 200 stocks and the group comprising the 1001st to 2000th stocks ranked by size, and the average difference across size cohorts between the returns of value and growth stocks. In the effective asset mix regressions, fitted returns are generated from regressions on either six Wilshire Target Indexes; six portfolios from independent sorts by size (large, small) and book-to-market (growth, neutral, and value); or six portfolios from sorts by size (large, mid-, and small capitalization), and the conditional value composite variable (value, growth). The coefficients of the regressors are constrained to be nonnegative and to sum to one. In the cross-sectional regression approach the portfolio's benchmark return is the weighted average of the fitted returns of each stock held in the portfolio using beginning-of-quarter portfolio weights. Fitted returns are from a cross-sectional regression of individual stock returns over the quarter on indicator variables for stock size, beginning-of-quarter rank values of book-to-market, cash flow yield, dividend yield, earnings yield, sales-to-price ratio, past six-month return, and industry dummy variables.

entire history increases the precision of the estimated loadings. However, this tends to overfit the data, and as a result confounds managerial skill with the portfolio's exposures. To avoid overfitting, we exclude the quarter under evaluation when we estimate the predicted return. Both sets of results are reported in the tables.

As in Table 1, there is a wide range in abnormal returns estimated from the regression-based benchmarks. Average abnormal returns for the entire sample of managers over the full period (panel A of Table 6) vary from 3.67% to −1.97%, yielding a range of 5.64%.[22] Even when we narrow attention to the time-series regressions using factor-mimicking portfolios, mean abnormal returns are 2.64% and 3.67% on an out-of-sample basis. Similarly, the effective asset mix regressions produce out-of-sample mean abnormal returns between 1.48% and 3.09%. These differences stand out all the more because they are generated from models that closely resemble one another, are fitted over many quarters, and are averaged over numerous portfolios.

During the relatively short 1998:Q1–2000:Q1 subperiod, the difference in benchmark-adjusted returns across methods is even more acute, with the range rising to 14.31%.[23] Within the set of Sharpe (1992) style regressions, out-of-sample abnormal returns display a range of 5.20%.

The message from Table 6 is that abnormal return estimates are very sensitive to the choice of regressors. This is the case although the regressors tend to be highly correlated. For instance, in the effective asset mix regressions, the average pairwise correlation is 0.97 between the return series on the large-growth style indexes, and 0.88 for the large-value style indexes. The overall average pairwise correlation between the corresponding regressors in the style regressions is 92%.

The noise in active portfolio returns and their limited histories make it hard to discriminate between the procedures in terms of tracking error volatility.[24] Nonetheless, two procedures stand apart from the others in terms of their poor out-of-sample performance. The cross-sectional regression approach includes a variety of stock attributes. However, it generates the largest tracking error volatility of all the models for the full period (10.60%).[25] The tracking error volatility from the Fama-French (1993) three-factor model is 10.54%, which

---

[22] A test that the regression-based methods all produce the same mean abnormal return for the equally weighted portfolio of managers yields an F-statistic (with standard error corrected for clustering by calendar quarter) of 5.20 with a p-value of less than 1%.

[23] Note that the statistics reported in panel B use the nine quarterly abnormal returns for each managed portfolio over the subperiod. It is still the case, however, that each abnormal return is based on model parameters estimated over the portfolio's entire history (either in full, or omitting the quarter where performance is measured) for the full sample period.

[24] For example, the F-statistic is 1.25 (p-value of 0.30) for the hypothesis that the equally weighted portfolio of managers has the same out-of-sample tracking error volatility across the methods in Table 6.

[25] Some additional experiments suggest that several other specifications of the cross-sectional model yield even higher tracking error volatility. In particular, when the size variable is measured as the logarithm of market capitalization, the mean tracking error volatility rises to 13.04%.

is substantially higher than the results from other methods.[26] The volatility rises to 14.77% during the late 1990s. Not all is lost for the factor model, however. Revised factor-mimicking portfolios for size and the size-conditional composite value indicator knock the out-of-sample tracking volatility down to 8.33% for the overall period and 9.08% for the 1998–2000:Q1 epoch.

Volatilities of tracking errors from the regresssion-based benchmarking models in Table 6 are roughly comparable in magnitude to those from the characteristic-based models in Table 1. Value-weighted control portfolios from the size-dependent sort on the composite value indicator generate mean tracking error volatilities of 8.71% for the full period. This is in line with the better regression-based models: the factor model using mimicking portfolios based on size and the composite value score produces an out-of-sample mean tracking error volatility of 8.33% for the full period. However, net-of-benchmark returns from the regression models tend to be larger in absolute terms.

Table 7 compares the performance of the regression-based benchmarks within each of four investment styles. To minimize clutter, we report results only for the benchmarks that leave out the evaluation quarter from the estimation period. The results reinforce the key findings from the overall sample in the previous table. First, there is a large range across methods in mean abnormal returns, even when we limit attention to homogeneous sets of portfolios. Within the category of large-value managers, for example, the range across methods in average performance levels is 4.88%. During the 1998–2000 subperiod, the range is even more breathtaking (10.52%). The dispersion in abnormal returns across procedures, as well as the generally large absolute magnitude of mean abnormal returns, do not inspire confidence in the regression-based benchmarks. In contrast, the characteristic-based benchmarks in Table 2 produce abnormal returns that are closer to zero, even during the 1998–2000 epoch.

Second, the tracking error volatilities of the large-capitalization portfolios are closely bunched with the exception of the Fama-French (1993) factor model and the cross-sectional regression. In the case of large-value portfolios, for example, the tracking volatilities from these two procedures exceed 8.50%, while volatilities from the other methods are much lower (less than 6%).

Third, tracking volatilities from the regression-based benchmarks and the characteristic-based benchmarks are generally comparable in the case of large-cap portfolios. When the characteristic-based methods are applied to the combined large growth and value sample, the size, composite value indicator approach is associated with the lowest tracking error volatility (6.89%). In the same sample, the best performing benchmark from the regression models has a mean tracking volatility of 6.81%. However, the estimated abnormal returns are much more different: 0.54% from the characteristic-based approach as opposed

[26] The results from the Fama-French (1993) three-factor model are also more sensitive to the omission of the evaluation quarter. This finding suggests that the estimated three-factor model loadings for managed portfolios are not very stable over time. The average in-sample tracking error volatility is 7.94%, compared to the mean out-of-sample volatility of 10.54%. The differences are less pronounced for the other methods.

Table 7
Performance (in percentage per year) of managed portfolios using alternative regression-based benchmarks, classified by investment style

(A) Large-capitalization portfolios

	Large-growth portfolios				Large value portfolios					
	Abnormal return		Tracking error volatility		Abnormal return		Tracking error volatility		Large-cap portfolios	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean abnormal return	Mean tracking error volatility
					Panel A1: Full period, 1989:Q1–2001:Q4					
Fama-French three factors	5.54*	5.02	8.83	7.68	−0.17	−0.01	8.67	7.64	3.17*	8.76
Market, size, value composite factors	3.41*	2.64	7.86	7.08	3.48*	2.90	5.33	4.74	3.44*	6.81
Effective asset mix regressions:										
Wilshire indexes	2.28*	1.55	8.01	7.10	3.30	3.04	5.98	5.64	2.71*	7.16
Independent sort size, BM portfolios	3.23*	2.05	7.86	6.08	0.40	−0.21	5.98	5.34	2.05*	7.35
Size, value composite portfolios	3.41	2.64	7.86	7.08	2.00	1.58	5.97	5.34	2.16	7.43
Cross-sectional regression	−2.28	−2.51	11.04	10.72	−1.40*	−1.66	8.70	6.96	−1.92	10.07
					Panel A2: 1998:Q1–2000:Q1					
Fama-French three factors	10.08	5.73	12.55	10.80	−5.59	−5.61	13.18	12.26	2.64	12.85
Market, size, value composite factors	5.26	3.65	9.51	7.92	3.47	2.65	5.61	5.06	4.41	7.66
Effective asset mix regressions:										
Wilshire indexes	10.39	5.33	10.12	8.00	4.93	5.10	6.69	5.88	7.80	8.49
Independent sort size, BM portfolios	11.03	6.13	10.82	9.32	−1.59	−1.96	7.00	5.90	5.04	9.01
Size, value composite portfolios	7.20	5.01	10.65	8.82	2.56	1.97	6.81	5.56	5.00	8.83
Cross-sectional regression	−19.12	−16.49	17.01	16.22	2.66	4.30	10.94	8.12	−8.78	14.13

(B) Small-capitalization portfolios

Panel B1: Full period, 1989:Q1–2001:Q4										
Fama-French three factors	5.49	4.87	20.16	21.48	−1.42	−2.28	9.30	6.86	2.03	14.73
Market, size, value composite factors	2.98	1.75	14.50	14.62	4.63*	4.23	9.37	7.88	3.80	11.93
Effective asset mix regressions:										
Wilshire indexes	3.34	2.52	15.30	14.88	5.16*	4.27	9.59	7.84	4.25	12.45
Independent sort size, BM portfolios	3.74*	4.42	12.94	12.28	−2.05	−2.88	8.73	7.88	0.84	10.84
Size, value composite portfolios	0.92	2.06	13.83	13.38	3.45*	3.31	8.91	7.34	2.18	11.37
Cross-sectional regression	−5.47	−3.77	14.94	14.00	1.39	1.17	8.44	7.54	−2.04	11.69
Panel B2: 1998:Q1–2000:Q1										
Fama-French three factors	22.29	20.81	35.15	36.34	−1.17	−4.37	10.10	8.24	8.43	20.34
Market, size, value composite factors	5.04	3.95	18.22	16.58	2.64	1.79	9.01	8.40	3.62	12.78
Effective asset mix regressions:										
Wilshire indexes	21.55	19.85	19.68	19.08	4.33	2.73	9.37	9.46	11.37	13.59
Independent sort size, BM portfolios	10.38	9.95	15.96	14.82	−7.81	−9.83	8.60	8.10	−0.37	11.61
Size, value composite portfolios	1.55	4.46	18.12	15.54	2.63	1.94	9.45	8.78	2.19	13.00
Cross-sectional regression	−12.53	−12.47	15.99	16.80	−0.68	−0.66	8.00	5.90	−5.52	11.27

Each quarter, a managed portfolio's benchmark return is the fitted value from one of a variety of regression models. A managed portfolio's performance is measured as its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4), and during 1998:Q1–2000:Q1. A portfolio's mean abnormal return is its annualized geometric mean return minus the annualized geometric mean return on the benchmark. A portfolio's tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the portfolio's return and the benchmark's return. For each performance measure, the arithmetic mean and median are provided over the cross-section of 199 managed portfolios in the sample period. An asterisk associated with the mean abnormal return over the full sample period denotes that the time-series mean of the quarterly equally weighted average return in excess of the benchmark across all available managed portfolios is at least two standard errors away from zero. Results are provided for large-capitalization (growth or value) portfolios in panel A, and for small-capitalization (growth or value) portfolios in panel B. A portfolio's investment style is based on its self-reported style where available, and otherwise based on its scaled rank by size and composite value indicator. In panels A and B, the regression uses all quarters over the full period 1989–2001, or excludes the current quarter to estimate coefficients. The estimated slope coefficients, along with the realized values of the regressors in the current quarter, are used to generate the fitted value (any intercept term is suppressed). In the Fama-French (1993) three-factor model, the regressors are the market excess return and returns on mimicking portfolios for size and book-to-market, *SMB* and *HML*. In the market, size, and value composite factor model, the regressors are the market excess return, the difference between the value-weighted return on the largest 200 stocks and on the group comprising the 1001st to 2000th stock ranked by size, and the average difference across size cohorts between the returns of value and growth stocks. In the effective asset mix regressions, fitted returns are generated from regressions on either six Wilshire Target Indexes; six portfolios from independent sorts by size (small, large), and book-to-market (growth, neutral, and value); or six portfolios from sorts by size (small, mid-, and large), and the conditional value composite variable (value, growth). The coefficients of the regressors are constrained to be nonnegative and to sum to one. In the cross-sectional regression approach, the portfolio's benchmark return is the weighted average of the fitted returns of each stock held in the portfolio using beginning-of-quarter portfolio weights. Fitted returns are from a cross-sectional regression of individual stock returns over the quarter on indicator variables for stock size, beginning-of-quarter rank values of book-to-market, cash flow yield, dividend yield, earnings yield, sales-to-price ratio, past six-month return, and industry dummy variables.

to 3.44% from the factor regression model. Insofar as the outsized levels of performance from the regression models are hard to reconcile with intuition, the advantage seems to go to the characteristic-based methods.

Fourth, the volatile behavior of small-stock returns generally prevents clear-cut distinctions between the methods. Nonetheless, the Fama-French (1993) factor model still turns in the poorest showing with respect to tracking error volatility. In the combined set of small-capitalization growth and value portfolios, the mean tracking volatility is 14.73% for the overall period.[27] In contrast, for the same set of portfolios, the revised factor model based on size and the composite value score features a mean tracking volatility of 11.93%.

5. Interpreting the Evidence from Regression-based Methods

To discriminate more sharply between the benchmarks derived from regression models, we apply them to the Russell indexes. Table 8 reports on the results.

Passive indexes should yield no indication of performance. Further, the regression models are estimated over the full 13-year history, so there should be little margin for disagreement. It is thus startling that the regression-based benchmarks produce sizeable spreads in abnormal returns. In the case of the Russell 1000 growth index, the range across methods exceeds 3% and in the case of the Russell 1000 value index, the range is above 2%. The ranges are substantially larger for the small-stock indexes. For the Russell 2000 value index, for instance, one method reports an impressive mean net-of-benchmark return of 3.50%, while another equally sensible method suggests that performance is a disastrous −3.30%.

The Fama-French (1993) factor model is at the core of academic research on investment performance. Tracking error volatility from this approach, however, is exceptionally high. Averaged across all the indexes, the mean out-of-sample tracking volatility from the Fama-French (1993) three-factor model is 4.99%. An adjustment to the factor model that uses a comprehensive measure of value/growth within each size cohort is more successful: out-of-sample tracking error volatility is cut substantially to 3.67% on average.[28] The cross-sectional regression approach is popular in academic research and widely used by practitioners as well. This method also shows subpar performance, generating tracking volatility of 4.85% on average across the Russell indexes.

---

[27] While the cross-sectional regression method has poor tracking performance for large-stock portfolios, it performs on par with the other methods for small-stock portfolios. The cross-section comprises many more small stocks than large, and the variation in returns and attributes is more pronounced for small stocks than for large. As a result, the regression model tends to accommodate the behavior of small stocks, everything else being equal. Consequently, fitted returns from the cross-sectional regression have an easier time tracking small stocks.

[28] In additional experiments, we verify that the bulk of the improvement stems from the modified value factor-mimicking portfolio rather than the modified size factor. We do this by fitting a factor model that includes the market, the conventional size factor *SML* used by Fama and French (1993), and the value factor based on the composite value indicator. This approach yields a lower tracking error volatility for six of the eight Russell indexes compared to a model that combines the market factor, the conventional Fama-French value factor *HML*, and the modified size factor.

Table 8
Performance (in percentage per year) of Russell indexes using alternative regression-based benchmarks

						Effective asset mix regressions with						
		Fama-French three-factor model		Market, size, value composite factor model		Wilshire indexes		Independent sort size, BM portfolios		Size, value composite portfolios		Cross-sectional regression on attributes
		All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	
						Panel A: Full period, 1989:Q1–2001:Q4						
Top 200 Growth	Abnormal return	0.71	0.62	−0.55	−0.60	−1.68	−1.70	−1.01	−1.01	−0.77	−0.72	0.95
	Tracking error	4.66	5.16	3.38	3.60	3.30	3.32	3.74	3.74	2.74	3.04	6.54
Top 200 Value	Abnormal return	0.26	0.28	0.86	0.88	0.65	0.65	−0.97	−1.06	−0.20	−0.05	−0.33
	Tracking error	3.60	3.86	3.16	3.42	3.76	3.86	3.54	3.78	2.76	3.00	5.66
Mid-cap growth	Abnormal return	0.05	0.11	−0.29	−0.16	−0.43	−0.39	2.05	2.09	−2.61*	−2.42	1.98*
	Tracking error	5.94	6.60	5.12	5.70	9.28	9.76	7.72	7.96	4.04	4.68	3.96
Mid-cap value	Abnormal return	−0.10	−0.20	2.11*	2.11*	0.81	1.00	−1.35	−1.41	0.30	0.43	1.59
	Tracking error	5.34	5.80	3.06	3.40	2.58	2.90	3.36	3.80	2.20	2.46	3.58
R1000 Growth	Abnormal return	0.25	0.16	−0.81	−0.86	−1.98	−2.00	−1.00	−1.00	−1.44	−1.29	1.33
	Tracking error	3.68	4.10	2.22	2.38	3.04	3.10	3.54	3.64	2.54	3.04	5.42
R1000 Value	Abnormal return	0.00	−0.02	1.18	1.17	0.56	0.63	−0.94	−0.99	−0.12	0.01	0.33
	Tracking error	3.50	3.80	2.34	2.56	2.72	2.96	2.34	2.50	2.18	2.42	4.64
R2000 Growth	Abnormal return	−4.74*	−4.85*	−3.76*	−3.76*	−3.30	−3.19	−0.64	−0.70	−3.17*	−2.74	−2.76
	Tracking error	3.64	3.88	3.40	4.00	6.94	7.20	2.86	3.00	4.42	5.42	5.70
R2000 Value	Abnormal return	−0.02	−0.22	3.50*	3.36*	0.95	1.01	−3.18*	−3.30*	2.27*	2.28	2.00*
	Tracking error	6.00	6.70	3.88	4.28	5.56	5.86	3.86	4.16	3.50	3.88	3.30
Average	Abnormal return	−0.45	−0.52	0.28	0.27	−0.55	−0.50	−0.88	−0.92	−0.72	−0.56	0.64
	Absolute abnormal return	0.77	0.81	1.63	1.61	1.30	1.32	1.39	1.44	1.36	1.24	1.41
	Tracking error volatility	4.55	4.99	4.22	3.67	4.65	4.87	3.87	4.07	3.05	3.49	4.85
						Panel B: 1998:Q1–2000:Q1						
Top 200 Growth	Abnormal return	0.96	0.20	−3.81	−4.16	−1.86	−1.86	3.53	3.53	1.38	1.57	−10.39
	Tracking error	5.06	6.24	3.98	4.48	2.00	2.00	3.50	3.50	2.74	2.92	11.52
Top 200 Value	Abnormal return	0.73	0.95	2.42	2.41	6.97	7.25	3.13	3.52	2.14	2.29	10.52
	Tracking error	3.78	4.14	3.56	3.96	2.58	2.60	3.40	3.76	3.26	3.46	8.64

(*Continued overleaf*)

Table 8
(*Continued*)

		Fama-French three-factor model		Market, size, value composite factor model		Effective asset mix regressions with: Wilshire indexes		Independent sort size, BM portfolios		Size, value composite portfolios		Cross-sectional regression on attributes
		All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	All quarters	Exclude current quarter	
Mid-cap growth	Abnormal return	3.38	4.10	0.59	0.79	17.48	18.39	13.05	13.48	−4.49	−3.37	2.16
	Tracking error	6.90	7.84	8.70	9.60	15.06	15.94	11.02	11.52	5.70	6.90	2.52
Mid-cap value	Abnormal return	−5.13	−5.35	−0.08	−0.38	2.54	2.92	−6.06	−6.53	0.50	0.93	3.06
	Tracking error	3.98	4.38	2.52	2.78	1.62	1.94	3.18	3.62	3.64	4.06	4.78
R1000 Growth	Abnormal return	0.97	0.43	−3.15	−3.39	−0.53	−0.60	4.45	4.62	0.00	0.96	−8.15
	Tracking error	3.72	4.66	2.38	2.68	2.40	2.48	3.70	3.94	2.94	3.58	9.34
R1000 Value	Abnormal return	−1.44	−1.37	1.42	1.30	4.50	5.00	−0.31	−0.15	1.50	1.76	8.11
	Tracking error	3.28	3.62	2.60	2.92	2.22	2.42	2.88	3.10	2.64	2.88	7.38
R2000 Growth	Abnormal return	−6.42	−6.44	−4.43	−4.19	8.49	9.09	−2.76	−2.82	−3.40	−2.19	−5.44
	Tracking error	3.06	3.48	3.60	4.32	8.42	8.94	2.58	2.68	4.96	6.34	7.60
R2000 Value	Abnormal return	−4.25	−4.70	1.80	1.35	2.25	2.16	−10.05	−10.84	2.10	2.33	0.59
	Tracking error	4.64	5.06	4.50	5.04	5.04	5.42	3.60	4.18	5.66	6.20	4.32
Average	Abnormal return	−1.40	−1.52	−0.66	−0.78	4.98	5.29	0.62	0.60	−0.03	0.54	0.06
	Absolute abnormal return	2.91	2.94	2.21	2.25	5.58	5.91	5.42	5.69	1.94	1.93	6.05
	Tracking error volatility	4.30	4.93	3.98	4.47	4.92	5.22	4.23	4.54	3.94	4.54	7.01

Each quarter, the benchmark return for a Russell style index is calculated as the fitted value from a regression model, using one of several procedures. For each index, statistics are provided for its mean abnormal return and tracking error volatility over the entire sample period (1989:Q1–2001:Q4), and during 1998:Q1–2000:Q1. The mean abnormal return is the annualized geometric mean return on the index minus the annualized geometric mean return on the benchmark return. An asterisk denotes that the mean abnormal return is more than two time-series standard errors away from zero. The tracking error volatility is the annualized standard deviation of the time series of quarterly differences between the index return and the benchmark's return. The indexes are the Russell top 200 (value and growth), the Russell mid-cap (value and growth), the Russell 1000 (value and growth), and the Russell 2000 (value and growth). The performance measures under each procedure are also averaged across the eight indexes and reported at the bottom of each panel. In Panels A and B, the regression uses all quarters over the full period 1989–2001, or excludes the current quarter to estimate coefficients. The estimated slope coefficients, along with the realized values of the regressors in the current quarter, are used to generate the fitted value (any intercept term is suppressed). In the Fama-French (1993) three-factor model the regressors are the market excess return and returns on mimicking portfolios for size and book-to-market, *SMB* and *HML*. In the market, size, and composite value indicator factor model, the regressors are the market excess return, the difference between the value-weighted return on the largest 200 stocks and on the group comprising the 1001st to 2000th stock ranked by size, and the average difference across size cohorts between the returns of value and growth stocks. In the effective asset mix regressions, fitted returns are generated from regressions on either six Wilshire Target Indexes; six portfolios from independent sorts by size (large, small) and book-to-market (growth, neutral, and value); or six portfolios from sorts by size (large, mid-, and small) and the conditional value composite variable (value, growth). The coefficients of the regressors are constrained to be nonnegative and to sum to one. In the cross-sectional regression approach, the portfolio's benchmark return is the weighted average of the fitted returns of each stock held in the portfolio using beginning-of-quarter portfolio weights. Fitted returns are from a cross-sectional regression of individual stock returns over the quarter on indicator variables for stock size, beginning-of-quarter rank values of book-to-market, cash flow yield, dividend yield, earnings yield, sales-to-price ratio, past six-month return, and industry dummy variables.

Sharpe (1992) style regressions generally fare well. In asset mix regressions using the style portfolios underlying *SMB* and *HML*, out-of-sample standard deviations of abnormal returns average 4.07%. Regressions that use the size, composite value reference portfolios produce mean tracking error volatilities of 3.49%. The latter model generates the lowest tracking error volatility for six out of the eight indexes.[29]

The relatively well-behaved nature of the passive indexes provides leverage in discriminating between the models' performance. From this standpoint, the characteristic-based approach has a slight edge over the regression-based approach. Of the characteristic-based models in Table 3, the best performing benchmarking model uses reference portfolios from sorts by size and then within each size group by the composite value indicator. Tracking error volatilities from this approach average 3.01%. Characteristic matching procedures predict returns using stock attributes that are known at the beginning of the quarter. The most direct analogs are thus the regression procedures where the estimation period is divorced from the evaluation period. Among the approaches in Table 8, the model with the best out-of-sample performance is the effective asset mix regression using the same reference portfolios. The resulting standard deviation of tracking errors is higher at 3.49%.

6. Benchmark Choice and Portfolio Performance

An investor, financial advisor, or money manager is concerned with how an individual portfolio performs. Performance can be judged against a variety of benchmarks, each of which is logically compelling. The issue, therefore, is how closely the procedures agree as to the magnitude of performance when they are applied to the same portfolio. Additionally, since there is no clear-cut choice as to the appropriate horizon over which performance is evaluated, we explore the differences across methods at different time intervals.

Degeorge, Patel, and Zeckhauser (1999) argue that many economic agents care about thresholds, and attach importance to specific demarcations (such as positive versus negative earnings). For investors, the benchmark's return serves as an important reference point. In such circumstances, a portfolio manager is especially concerned with crossing that threshold and reporting positive performance. Accordingly, we also see whether the methods tend to agree at least on whether there is over- or underperformance.

---

[29] The style regressions generally do poorly at tracking small-cap portfolios. One objection is that the regressors in the model include large- and mid-cap portfolios, which are not relevant to a small-cap style. Since the regressors are correlated, the regression may try to allocate some weight to equity styles that the portfolio is not oriented toward, thus clouding tracking ability. When we reestimate the regressions using only small-cap style indexes as regressors, the tracking error volatilities are reduced for the Russell 2000 growth index but not for the Russell 2000 value index. For example, in the Sharpe (1992) regression of the Russell 2000 growth index on small-cap style portfolios from sorts on size and the composite value indicator, the out-of-sample tracking volatility is 1.74%. Applied to the Russell 2000 value index, the same model has a tracking volatility of 4.72%. Since we choose to apply a uniform model to all portfolios, we do not pursue this modification of the Sharpe approach.

In Table 9, we provide comparisons across all the methods considered above. Given their prominence in the research literature, we also report more targeted contrasts for reference portfolios from independent sorts and the Fama-French (1993) three-factor model. We compare them to each other, and with either characteristics-based or regression-based alternatives. The independent sort portfolios are paired with our size, value composite benchmarks, and the Fama-French (1993) three-factor model is compared to the model based on the market and our size, value composite factors. To eliminate clutter, we give results only for value-weighted portfolios, and regressions fitted over the full sample.

The methods are compared with respect to average abnormal returns over the portfolio's entire history (panel A of Table 9), or with annual and quarterly abnormal returns (panels B and C of Table 9). In panel A, a portfolio's average abnormal return over its full history is the difference between the geometric mean return on the portfolio and on the benchmark. This is calculated for each benchmarking method under comparison. We take all pairwise differences in abnormal returns across methods and count how many absolute differences exceed a threshold level (2.5% or 5% per year) out of the total number of possible comparisons. The relative frequency is then averaged across portfolios. In addition, we calculate the fraction of portfolios where the methods agree on the sign of the abnormal return (estimated abnormal returns are either all positive or all negative). Panels B and C perform the same calculations using abnormal returns measured over each calendar year or each quarter, and average the results across portfolios and across time.

Since our benchmarking procedures share the premise that size and value/growth orientation are the key drivers of average returns, the presumption is that they should agree on the sign of a portfolio's performance, if not the magnitude of the performance. The results in Table 9 suggest that such consensus happens less frequently than might be hoped. Comparing across all our methods, in the overall sample period there is at least one disagreement about the sign of a portfolio's abnormal return in 79.40% of the portfolios (or in only 20.60% of the cases do the methods agree on the sign of abnormal returns). Even when we single out methods that are highly similar on the surface, the frequency of disagreement is considerable. Average abnormal returns estimated from factor regression models, but using different approximations for the size and value factors, have different signs in 25.16% of the cases.

More specifically, the methods are likely to produce mean abnormal returns that deviate notably from one another. Across all our methods, absolute differences in excess of 2.5% per year occur with a frequency of 39.79%, and absolute differences above 5% per year occur in 16.76% of the cases. As noted in the introduction, the two methods most extensively used in research (independently sorted size and book-to-market control portfolios, and the Fama-French 1993 factor model) generate average abnormal returns that deviate by as much as 2.5% with 43.22% frequency. The likelihood of a large difference across methods rises during the turbulent 1998–2000 subperiod.

Table 9
Frequency of differences in measured abnormal return across benchmarks

Panel A: Average abnormal annualized return

	Percentage of comparisons yielding					
		Absolute differences above			Absolute differences above	
Methods compared	Different signs	2.5%	5%	Different signs	2.5%	5%
	Full period, 1989:Q1–2001:Q4			1998:Q1–2000:Q1 subperiod		
All methods	0.7940	0.3979	0.1676	0.9000	0.7025	0.4845
Independent size, BM portfolios and Fama-French three factors	0.2462	0.4322	0.1407	0.1750	0.6438	0.4375
Independent size, BM and size, value composite portfolios	0.2059	0.2549	0.0229	0.3038	0.8423	0.6077
Fama-French three factors and market, size, value composite factors	0.2516	0.4444	0.1895	0.2731	0.6385	0.3731

Panel B: Annual abnormal return

	Percentage of comparisons yielding					
		Absolute differences above			Absolute differences above	
Methods compared	Different signs	2.5%	5%	Different signs	2.5%	5%
	Full period, 1989:Q1–2001:Q4			1998:Q1–2000:Q1 subperiod		
All methods	0.8144	0.6232	0.4028	0.8681	0.7250	0.5278
Independent size, BM portfolios and Fama-French three factors	0.2449	0.6565	0.4136	0.3453	0.7818	0.6235
Independent size, BM and size, value composite portfolios	0.3600	0.5512	0.3170	0.4460	0.7482	0.5228
Fama-French three factors and market, size, value composite factors	0.5968	0.5943	0.3591	0.5755	0.7050	0.4772

(*Continued overleaf*)

Table 9
(*Continued*)

Panel C: Quarterly abnormal return

Methods compared	Percentage of comparisons yielding: Different signs	Absolute differences above 1%	Absolute differences above 3%	Different signs	Absolute differences above 1%	Absolute differences above 3%
	Full period, 1989:Q1–2001:Q4			1998:Q1–2000:Q1 subperiod		
All methods	0.7808	0.6135	0.2620	0.8403	0.6913	0.3357
Independent size, BM portfolios and Fama-French three factors	0.2898	0.6953	0.3502	0.3285	0.8389	0.5153
Independent size, BM and size, value composite portfolios	0.1626	0.5018	0.1505	0.2264	0.6840	0.2486
Fama-French three factors and market, size, value composite factors	0.1587	0.4574	0.1564	0.2042	0.5910	0.2493

A managed portfolio's performance is measured under different benchmarking methods. In panel A, benchmarks are assessed based on the active portfolio's average abnormal return (the difference between the portfolio's geometric mean return and the benchmark's geometric mean return). The results tabulate the number of portfolios (out of a total of 199 active managers in the sample) where the methods yield mean abnormal return estimates that differ in sign. The fraction of pairwise comparisons across methods for a portfolio where the absolute difference in estimated abnormal annualized return exceeds either 2.5% or 5% per year is also calculated for each portfolio, and averaged across portfolios. In panels B and C, abnormal returns (portfolio return minus benchmark return) are measured for each full calendar year over a portfolio's history or each quarter, respectively, under each method. For each portfolio, the fraction of years or quarters where methods differ on the sign of the abnormal return is calculated, as well as the fraction of pairwise comparisons across methods where the absolute difference between abnormal returns exceeds a threshold level. The fractions are then averaged across all portfolios in the sample. The threshold levels are 2.5% and 5% per year, or 1% and 3% per quarter. Results are provided for all active portfolios over the full period 1989:Q1–2001:Q4 and the 1998:Q1–2000:Q1 subperiod. Two sets of benchmarking procedures are applied to every portfolio in the sample. Under the attribute-matched procedures, each stock held in a managed portfolio at the beginning of a quarter is matched against a reference portfolio. Under the independent sorting procedure, there are 25 control portfolios from the intersection of independent sorts by size (market value of equity) and BM (the ratio of book value of common equity to the market value of common equity). Under the size, within-size BM sort procedure, there are 28 control portfolios from sorts first by size, and then within each size category, by BM. In the size, value composite approach, a stock is given an overall ranking, conditional on its size group, based on book-to-market, dividend yield, cash flow yield, average earnings yield (based on the past year's net income, forecasted next year earnings, and forecasted two-year ahead earnings), and sales-to-price ratio. In these methods, the component stocks in a control portfolio are refreshed once a year at the end of June. In the quarterly size, within-size BM approach, the component stocks are refreshed at the beginning of each quarter. The return on a control portfolio is either the equally weighted or value-weighted average of the buy-and-hold returns on the component stocks. Each managed portfolio is also paired with a Russell style index depending on its investment style based on its self-reported style where available, and otherwise based on its scaled rank on size and conditional value composite indicator. For the regression-based benchmarking procedures, in each quarter the return on the benchmark is the fitted value from a regression of quarterly managed portfolio returns on different regressors. The regression uses all quarters over the full period 1989–2001, or excludes the current quarter, to estimate coefficients. The estimated slope coefficients, along with the realized values of the regressors in the current quarter, are used to generate the fitted value (any intercept term is suppressed). In the Fama-French (1993) three-factor model, the regressors are the market excess return and returns on mimicking portfolios for size and book-to-market, *SMB* and *HML*. In the market, size, and value composite factor model, the regressors are the market excess return, the difference between the value-weighted return on the largest 200 stocks and on the group comprising the 1001st to 2000th stock ranked by size, and the difference between the returns of value and growth stocks. In the effective asset mix regressions, fitted returns are generated from regressions on either six Wilshire Target Indexes; six portfolios from independent sorts by size, BM; or six portfolios from sorts by size and the conditional value composite variable. The coefficients of the regressors are constrained to be nonnegative and to sum to one. In the cross-sectional regression approach, the benchmark return on a portfolio is the weighted average of the fitted returns of each stock held in the portfolio using beginning-of-quarter portfolio weights. Fitted returns are from a cross-sectional regression of individual stock returns over the quarter on beginning-of-quarter values of stocks' size, book-to-market, cash flow yield, dividend yield, earnings yield, sales-to-price ratio, past six-month return, and industry dummy variables.

Given the emphasis on performance, a few quarters of poor results can sour relations between a money manager and clients. Panel B of Table 9 looks at the chances of disagreement, as well as the magnitude of differences across procedures in terms of year-by-year abnormal returns. Given the higher volatility of annual observations rather than full-history averages, the divergences across procedures are starker in panel B. In an average year, the methods agree on the direction of performance with a frequency of only 18.56%, and the chances of encountering absolute differences of above 5% in abnormal returns is 40.28%. Contrasting reference portfolios from independent sorts versus two-way within-group sorts, the frequency of absolute differences in excess of 5% is 31.70%. These results suggest that the choice of benchmarking procedure can make or break a money manager's reputation.

The evidence from yearly abnormal returns in panel B of Table 9 throws up a red flag about snap judgments regarding performance over short horizons. In practice, however, even a year is considered to be a long time and manager performance is often scrutinized over shorter intervals. Panel C examines the frequency and magnitudes of disagreements across methods with respect to quarterly abnormal returns. Note that the quarterly abnormal returns are roughly half as volatile as the annual series. Nevertheless, the differences across benchmarks yield a wide range of verdicts on performance. Across all methods, divergences in excess of 1% per quarter occur with a frequency of 61.35% over the full sample period, with the incidence growing to 69.13% in the 1998–2000 subperiod. Comparing the results from panel A to those from panel C in Table 9 offers the clear lesson that viewpoints about short-term performance rest on slippery footing. Averages over longer horizons, while still prone to a wide range in benchmark estimates, may yield a clearer assessment of individual manager performance.

7. Summary and Conclusions

Professional money managers invest large amounts of equity assets on behalf of pension plan sponsors, foundations, and individuals. In turn, clients are quick to hire and fire money managers on the basis of benchmarking metrics that aim to identify precisely which managers have beaten and are expected to beat the yardsticks. A large body of academic and practitioner research has extended the traditional Capital Asset Pricing Model and developed a broad array of methods to provide such benchmarks. Many of these methods, at first glance, appear to be slight tweaks of a common methodological approach based on size and value/growth as the main factors in the cross-section of returns. On the surface, then, it seems that the methods should all deliver more or less the same assessment about the level of manager performance.

Our analysis of a detailed dataset on money manager performance suggests that this is not the case. We use several variants of matched-characteristic reference portfolios and time-series return regressions to check for performance.

Estimated abnormal returns display large variation across procedures. For the sample of investment managers following a large-growth style, for instance, the range in mean abnormal returns across characteristic-based benchmarking methods is 3.79% and 7.82% across regression-based methods. The corresponding range across methods for large-value managers is 1.73% and 4.88%. Divergences across the methods in measured levels of performance are dramatically amplified during the overheated market conditions of 1998:Q1–2000:Q1. For the characteristic-based methods, the spread in mean abnormal returns of large-growth portfolios is 9.33% and across the regression-based methods, it is 30.15%. Applied to large-value portfolios, characteristic-based methods produce a range of 6.97% and regression-based methods generate spreads of 10.52%. These stark differences arise even though all the methods draw on the same premise that size and value/growth are the key drivers of stocks' average returns. As well, the methods are applied over a relatively extended period and averaged across numerous active portfolios.

Put another way, different methods applied to the same portfolio manager can produce abnormal returns that disagree dramatically with respect to sign and magnitude. In practice, managers are hired and fired on the basis of performance over relatively short horizons. In an average year, our full set of benchmarking methods agree on the sign of abnormal return with a frequency of only 18.56%. When the methods are compared in terms of the level of estimated abnormal returns in an average year, differences across methods in excess of 5% per year occur in 40.28% of the portfolios.

The import of these findings is that they suggest the following scenario can frequently occur. Suppose a client specifies a benchmark for an asset manager that correctly corresponds to the manager's style. Relative to this yardstick, the manager could outperform, perhaps even by a statistically significant margin. The client, however, could have selected another benchmark that is just as legitimate for the manager's style. Our results suggest that it would not be surprising to find that the same manager, without any change in behavior, underperforms the alternative yardstick. The frailty of inferences to the choice of benchmarking procedure, if not recognized, can impose substantial real costs. The process of terminating a manager whose performance is deemed to be unsatisfactory consumes resources. These expenses come in the form of hiring a transition manager, liquidating the portfolio, and the costs of searching for a replacement.

Our results let us assess the performance of benchmarking methods that have been extensively used in academic research and investment practice. Of these, the leading procedures that are widely used in academic research—characteristic-matched portfolios based on independent sorts by size and book-to-market, the three-factor time-series model with mimicking portfolios for size and book-to-market, and cross-sectional regressions of returns on a variety of predictors—have disappointing performance. They have poor ability to track the returns of both active and passive portfolios. As well, they are frequently

associated with implausible levels of over- or underperformance. Reference portfolios from independent sorts by size and book-to-market produce average abnormal returns of 5.31% when applied to the passive Russell 1000 growth index during the 1998:Q1–2000:Q1 period, and indicate performance of -5.64% for the Russell 1000 value index over the same period. For the same unmanaged indexes, the cross-sectional regression approach reports performance levels of −8.15% and 8.11%. We trace the poor showing of these methods to the underlying drawbacks in the design of the benchmarks.

Importantly, we discuss simple alterations that improve the performance of the benchmarking methods. Two-way within-group sorts by size and value/growth reflect more accurately the investment domains of equity asset managers. A comprehensive measure that takes other variables beyond book-to-market equity into account also matches portfolios' value/growth orientations better. More generally, benchmarks that are aimed at capturing the characteristics of active portfolios generally tend to have better performance than regression-based benchmarks. Capitalization-weighted control portfolios that match a managed portfolio's size and composite value indicator, when applied to the sample of active managers over the full period, produce a mean abnormal return of 0.78% and an average tracking error volatility of 8.71%. In comparison, the most widely applied benchmarking method in the academic research literature, the Fama-French (1993) three-factor regression model, generates a mean abnormal return of 2.64% and tracking error volatility of 10.54%. In the case of passive Russell indexes, the characteristic-matched procedure has an average tracking error volatility of 3.01%, compared to 4.99% for the three-factor model. Nonetheless, these findings may not dissuade a harsher observer from concluding that the general three-factor model, regardless of how the factors are measured, is insufficient for adequately capturing the return-generating process.

Our results are derived from a broad sample of managers, representing a variety of styles, and covering an extended period. Even so, the findings underscore the fuzziness surrounding judgments on investment performance in a standard context that is supposedly well understood. To sharpen this point, in practice performance tracking and attribution analysis often employs models with many factors over short periods. In light of the difficulty of filtering out managerial skill from investment style, verdicts on performance based on short horizons and overfitted models should be regarded with a healthy dose of skepticism.

References

Admati, A. R., and S. A. Ross. 1985. Measuring Investment Performance in a Rational Expectations Equilibrium Model. *Journal of Business* 58:1–26.

Barber, B. M., and J. D. Lyon. 1997. Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics. *Journal of Financial Economics* 43:341–72.

BARRA. 1990. *The United States Equity Handbook*. Berkeley, CA: BARRA.

Ben Dor, A., R. Jagannathan, and I. Meier. 2003. Understanding Mutual Fund and Hedge Fund Styles Using Return-based Style Analysis. *Journal of Investment Management* 1:97–137.

Carhart, M. M. 1997. On Persistence in Mutual Fund Performance. *Journal of Finance* 52:57–82.

Chan, L. K. C., H. L. Chen, and J. Lakonishok. 2002. On Mutual Fund Investment Styles. *Review of Financial Studies* 15:1407–37.

Chan, L. K. C., Y. Hamao, and J. Lakonishok. 1991. Fundamentals and Stock Returns in Japan. *Journal of Finance* 46:1749–64.

Chan, L. K. C., J. Karceski, and J. Lakonishok. 1999. On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model. *Review of Financial Studies* 12:937–74.

Chan, L. K. C., J. Lakonishok, and T. Sougiannis. 2001. The Stock Market Valuation of Research & Development Expenditures. *Journal of Finance* 56:2431–56.

Daniel, K., and S. Titman. 1997. Evidence on the Characteristics of Cross Sectional Variation in Stock Returns. *Journal of Finance* 52:1–33.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers. 1997. Measuring Mutual Fund Performance with Characteristic-based Benchmarks. *Journal of Finance* 52:1035–58.

Degeorge, F., J. Patel, and R. Zeckhauser. 1999. Earnings Management to Exceed Thresholds. *Journal of Business* 72:1–33.

Dybvig, P. H., and S. A. Ross. 1985. Differential Information and Performance Measurement Using a Security Market Line. *Journal of Finance* 40:383–99.

Fama, E. F. 1998. Market Efficiency, Long-Term Returns and Behavioral Finance. *Journal of Financial Economics* 49:283–306.

Fama, E. F., and K. R. French. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance* 46:427–66.

Fama, E. F., and K. R. French. 1993. Common Risk Factors in the Returns on Bonds and Stocks. *Journal of Financial Economics* 33:3–56.

Fama, E. F., and K. R. French. 1996. Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance* 51:55–87.

Fama, E. F., and K. R. French. 1997. Industry Costs of Equity. *Journal of Financial Economics* 43:153–93.

Ferson, W. E., and K. Khang. 2002. Conditional Performance Measurement Using Portfolio Weights: Evidence for Pension Funds. *Journal of Financial Economics* 65:249–82.

Ferson, W. E., and R. W. Schadt. 1996. Measuring Fund Strategy and Performance in Changing Economic Conditions. *Journal of Finance* 51:425–61.

Glosten, L. R., and R. Jagannathan. 1994. A Contingent Claim Approach to Performance Evaluation. *Journal of Empirical Finance* 1:133–60.

Goetzmann, W., J. Ingersoll, M. Spiegel, and I. Welch. 2007. Portfolio Performance Manipulation and Manipulation-Proof Performance Measures. *Review of Financial Studies* 20:1503–46.

Ikenberry, D., J. Lakonishok, and T. Vermaelen. 1995. Market Underreaction to Open Market Share Repurchases. *Journal of Financial Economics* 39:181–208.

Jegadeesh, N., and S. Titman. 1993. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance* 48:65–91.

Jensen, M. C. 1968. The Performance of Mutual Funds in the Period 1945–64. *Journal of Finance* 23:389–416.

Lakonishok, J., and I. Lee. 2001. Are Insider Trades Informative? *Review of Financial Studies* 14:79–111.

Lakonishok, J., A. Shleifer, and R. W. Vishny. 1994. Contrarian Investment, Extrapolation, and Risk. *Journal of Finance* 49:1541–78.

Lehmann, B. N., and D. M. Modest. 1987. Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons. *Journal of Finance* 42:233–65.

Lyon, J. D., B. M. Barber, and C. L. Tsai. 1999. Improved Methods for Tests of Long-Run Abnormal Stock Returns. *Journal of Finance* 54:165–201.

Merton, R. C. 1981. On Market Timing and Investment Performance. I. An Equilibrium Theory of Value for Market Forecasts. *Journal of Business* 54:363–406.

Mitchell, M. L., and E. Stafford. 2000. Managerial Decisions and Long-Term Stock Price Performance. *Journal of Business* 73:287–329.

Sharpe, W. F. 1992. Asset Allocation: Management Style and Performance Measurement. *Journal of Portfolio Management* 18:7–19.

Wermers, R. 2004. Is Money Really "Smart"? New Evidence on the Relation between Mutual Fund Flows, Manager Behavior, and Performance Persistence. Working Paper, University of Maryland.