ACCOUNTING & FINANCE

afaanz

Accounting and Finance 48 (2008) 761–781

Benchmarking benchmarks: measuring characteristic selectivity using portfolio holdings data

Kingsley Fong, David R. Gallagher, Adrian D. Lee

Australian School of Business, University of New South Wales, Sydney, 2052, Australia

Abstract

This study proposes methodological adjustments to the widely adopted performance benchmarking methodology of Daniel *et al*. (1997) as a means of improving the precision of alpha measurement for active equity fund managers. We achieve this by considering the monthly updating of characteristic benchmarks and to ensure neutrality to the Standard & Poor's/Australian Stock Exchange 300 index. Applying this benchmark to a representative sample of active Australian equity funds and simulated passive portfolios that mimic fund manager-style characteristics, we find statistically different and lower tracking error compared with using the standard characteristic benchmark methodology. We also find evidence that the modified benchmark statistically infers an alpha closer to zero compared with the standard benchmark methodology. Our findings suggest that improved specifications of characteristic benchmarks represent better methods in quantifying fund manager skill.

Key words: Equity fund benchmarking; Characteristic-based benchmarks

JEL classification: G12, G23

doi: 10.1111/j.1467-629x.2008.00263.x

This research was funded through an Australian Research Council Linkage Grant (LP0561160) involving Vanguard Investments Australia and Securities Industry Research Centre of Asia-Pacific. We are grateful for the helpful comments from a number of individuals, including an anonymous referee, Doug Foster, Eric Smith, Scott Lawrence and seminar participants at the 19th Australasian Banking & Finance Conference (2006), and the 20th University of Western Australia PhD Conference in Economics and Business (2007). The authors thank Vanguard Investments Australia for research support. The authors also thank the organizers of the 12th Annual Super Bowl of Indexing held in Scottsdale, Arizona, USA (2007), where this paper won the William F. Sharpe Award for best index-related research paper.

Received 20 September 2007; accepted 15 January 2008 by Robert Faff (Editor).

1. Introduction

Do active equity managers possess skill? Academics, investors, investment consultants and the financial press have been debating this issue as fees associated with actively managed funds should be justifiable. At the centre of this argument is an accurate benchmark to quantify fund manager skill. Although the literature has demonstrated the impossibility of constructing a perfect benchmark, improving benchmarking methods remains an important area of research.[1] In the case of stock portfolios, benchmark construction philosophy has evolved from market capitalization indexing to returns-based regression and holding-based methodologies that adjust for stock characteristics (or investment styles).

Daniel *et al*. (1997) propose an important method of incorporating style information in the use of characteristic-based benchmarks. Research findings based on such benchmarks has reopened the debate on the value of active management. For US mutual funds, Daniel *et al*. (1997), Wermers (2000) and Avramov and Wermers (2006), and, in the Australian context, Pinnuck (2003) and Gallagher and Looi (2006), find some evidence that active fund managers possess sufficient skill to earn returns to cover their costs, consistent with the Grossman and Stiglitz (1980) information equilibrium. This is in contrast to the literature, over a number of decades, documenting that active fund managers possess no skill when assessed on their aggregate net returns.[2]

The intuitive design and ease of implementation of the Daniel *et al*. (1997) benchmark has made it a popular choice by researchers with more granular portfolio information, such as portfolio holdings.[3] Chan *et al*. (2006) also show empirically that such benchmarking techniques work better in tracking passive styles than either the regression or independent sorting techniques of Fama and French (1996).

The present study proposes several modifications to the original Daniel *et al*. benchmark. First, we consider weighting characteristic benchmarks based on the composition of a commonly referenced broad-based index. This design results in a benchmark that assigns zero alpha to a pure capitalization-weighted index representative of the investable universe. Second, by using a monthly portfolio formation approach, we incorporate more timely characteristic information of a stock, compared with annual updating. Third, we use an overlapping

[1] For example, Chen and Knez (1996) show that there are an infinite set of admissible benchmarks which provide an infinite number of ranking orders. See also Roll (1977, 1978), Green (1986), Lehmann and Modest (1987), Kothari and Warner (2001) and Pástor and Stambaugh (2002a,b).

[2] See, for example, Jensen (1968), Malkiel (1995), Gruber (1996) and Ferson and Schadt (1996).

[3] See studies such as Coval and Moskowitz (2001) and Kacperczyk *et al*. (2005).

benchmark approach (similar to Jegadeesh and Titman, 1993, 2001) to better match the characteristic return of a stock.

Applying this benchmark to the portfolio holdings of active Australian equity managers from 1995 to 2002, we find a fall in tracking error volatility compared with the Pinnuck (2003) benchmark from 2.19 to 1.34 percent per year, This fall in tracking error volatility is statistically different using Levene's test for homogeneity of variances. Even when using a simpler non-overlapping benchmark, where characteristic benchmarks are monthly value-weighted and restricting to the Standard & Poor's/Australian Stock Exchange (S&P/ASX) 300 index, tracking error volatility is 1.43 per cent per annum and statistically different; hence, highlighting improvements to the benchmark with simple modifications.

The results are robust when we use the benchmarks to measure stock selection ability with respect to fund style with the overlapping benchmark approach. This shows lower tracking error volatility across fund styles, and is statistically different for three of the five fund styles (Growth at a Reasonable Price (GARP), Growth and Other) than the standard benchmark method. However, on both the aggregate and fund style level, the alpha of the overlapping benchmark is not statistically different to the standard benchmark, which suggests no difference in the quantitative ability of the benchmarks.

In further tests using portfolios that simulate passive investment manager fund styles (i.e. have *ex ante* zero alpha), we find that the modified benchmark achieves statistically lower and different tracking error based on Newey–West standard errors, and infers an alpha closer to zero compared with the standard Daniel *et al*. benchmark approach used by Pinnuck (2003). The overlapping benchmark is also superior to a market neutral characteristic benchmark, which does not control for size. In comparison to a benchmark without the monthly updating of benchmark characteristics, only lower tracking error is achieved. This suggests that the incremental improvement in the modified benchmark is in its focus on market neutrality. Our results are robust to out-of-sample testing in a period from July 2002 to December 2006. Taken together, the evidence indicates that more focused characteristic benchmarks within the investable domain enhance a performance analyst's quantification of fund manager skill.

The present paper is structured as follows. Section 2 outlines the data used and descriptive statistics. Section 3 describes our characteristic-based benchmark methodology, simulated portfolio methodology and benchmark statistical tests. Section 4 presents our empirical results and Section 5 concludes the paper.

2. Data

We collect month-end portfolio holdings data from the Portfolio Analytics Database (PAD). This database comprises the holdings of 38 active Australian equity fund managers. The database also contains self-reported fund styles

(GARP, Growth, Other, Style-Neutral and Value). Further details on this database are in Gallagher and Looi (2006). Monthly dilution-adjusted share returns and month-end market capitalizations are extracted from the Centre for Research in Finance (CRIF) Share Price and Price Relative (SPPR) database. Monthly returns of the S&P/ASX 300 Accumulation Index (S&P/ASX 300A) are sourced from Securities Industry Research Centre of Asia-Pacific. The Aspect Financial database is used for financial year-end book value (Aspect item ID 7010). Month-end weight compositions of the S&P/ASX 300 are sourced from Vanguard Investments Australia. Our sample period for PAD is from January 1995 to June 2002, although we extend our sample period to December 2006 using the other datasets for use in our passive portfolio simulations.

Table 1 presents the average monthly weight distribution of stocks held by our fund sample on a value-weighted basis sorted by size (MCAP), book-to-market ratio (BMC) and prior 1 year return (PR1YR) deciles. MCAP is the month-end market capitalization; BMC is the prior financial year book value over the month-end market capitalization; and PR1YR is the past 1 year return with 1 month lag. Panel A reports the distribution using the S&P/ASX 300 universe of stocks in benchmark formation and Panel B using the CRIF SPPR universe (i.e. all stocks listed on the ASX at any given time). There are approximately 260 stocks in the S&P/ASX 300 universe and 950 stocks in the CRIF SPPR universe at any given time that fulfils our data requirement above.[4]

Panel A shows that the funds underweight the largest 10 per cent of stocks in the S&P/ASX 300 by –1.76 per cent (MCAP decile 10). Over this period funds overweight stocks from deciles 7–9 or approximately the largest 60–120 stocks and underweight all other size deciles. This suggests that funds tend to concentrate their holdings in the top 200 stocks by market capitalization.[5]

Within weightings, BMC deciles 3–7 are overweight, suggesting that funds tend to hold moderate growth neutral stocks. Funds also favour stocks with high past returns as they overweight PR1YR deciles 6–9 although are weight neutral in the top decile.

Panel B shows that funds hold approximately 86 per cent of their portfolio value in the largest decile of stocks in the CRIF SPPR universe or in the 95 largest stocks. Overweighting in growth neutral (BMC deciles 3–6) and past winner stocks (except for the highest PR1YR decile) also occurs, similar to the evidence for the S&P/ASX 300 universe.

---

[4] Aside from being unable to account for initial public offering stock holdings due to lack of past returns data, our other limitations are the absence of book value data from the Aspect database for some stocks and omitting non-ordinary stocks that are not in the CRIF SPPR database.

[5] Another possible reason is that before April 2000, fund managers tracked the All Ordinaries Index; however, during April 2000, some funds benchmarked against the ASX 200 index while some tracked the ASX 300.

Table 1
Descriptive statistics

Panel A: Standard & Poor's/Australian Stock Exchange (S&P/ASX) 300 universe

MCAP	1	2	3	4	5	6	7	8	9	10
Fund weight	0.24	0.41	0.92	1.40	1.55	2.64	4.52	8.18	16.75	63.39
Fund: ASX 300	–0.16	–0.29	–0.09	–0.03	–0.41	–0.16	0.15	0.80	1.95	–1.76
BMC	1	2	3	4	5	6	7	8	9	10
Fund weight	5.05	12.72	18.04	20.87	15.88	11.40	7.47	4.21	2.88	1.49
Fund: ASX 300	–1.44	–2.04	0.05	2.67	2.45	1.92	0.04	–1.56	–1.38	–0.70
PR1YR	1	2	3	4	5	6	7	8	9	10
Fund weight	1.23	4.81	9.02	8.81	9.84	11.43	14.76	16.69	15.40	8.00
Fund: ASX 300	–0.55	–0.69	–0.33	–0.34	–0.18	0.25	0.84	0.36	0.65	0.00

Panel B: CRIF universe

MCAP	1	2	3	4	5	6	7	8	9	10
Fund weight	0.00	0.00	0.01	0.02	0.11	0.28	1.00	3.20	9.24	86.14
Fund: CRIF	–0.04	–0.08	–0.15	–0.23	–0.31	–0.47	–0.45	–0.06	0.26	1.53
BMC	1	2	3	4	5	6	7	8	9	10
Fund weight	5.09	15.49	26.11	20.37	15.72	9.67	4.19	2.04	1.13	0.20
Fund: CRIF	–3.16	–2.39	1.30	3.99	3.99	0.54	–1.68	–1.28	–0.94	–0.38
PR1YR	1	2	3	4	5	6	7	8	9	10
Fund weight	0.30	2.03	4.54	8.04	10.11	13.47	17.16	18.43	18.58	7.35
Fund: CRIF	–0.47	–0.84	–1.50	–1.04	–0.05	0.41	1.67	1.53	1.43	–1.13

At the end of each month from January 1995 to June 2002, stocks are ranked by their market capitalization (*MCAP*), book-to-market (*BMC*) and past 1 year return (*PR1YR*) independently into decile groups. 1 is the lowest decile group and 10 the highest. This table reports the monthly average weightings of the value-weighted Portfolio Analytics Database funds in stocks of different characteristic ranking, and their weighting differences against the CRIF Share Price and Price Relative (SPPR) and S&P/ASX 300 decomposed into these groupings. Panel A reports weighting decompositions in percentages for the S&P/ASX 300 universe and Panel B for stocks in the CRIF SPPR universe.

3. Research methodology

3.1. Characteristic-based benchmark methodologies

We use four characteristic-based benchmark methodologies. The standard benchmark, *Pinnuck*, and three alternatives: *Index*, *Broad* and *Overlap*. The standard benchmark methodology follows Pinnuck (2003). Every December month-end, stocks in the CRIF SPPR that fulfil data criteria for ranking by market capitalization, book-to-market and momentum are ranked by their current month-end market capitalization into five groups. Within each of these five groups, stocks are ranked and sorted by its book-to-market into four groups. Book-to-market is defined as the current year's book value divided by the current month-end market capitalization. Each group is then further sorted into

three groups by their past 11 month return, lagged 1 month. This is denoted as a 5/4/3 portfolio sort and results in 60 portfolios. The portfolios are held for 12 months based on value-weights by market capitalization. Note that value-weighting occurs at the beginning of the formation period and is fixed for the 12 month period unless a stock delists. If a stock delists, at the end of a month, the remaining stocks in that portfolio are reweighted by their past December-end market capitalization.

Our three alternative benchmarks are modifications to Pinnuck (2003) using stocks only in the S&P/ASX 300. *Index* uses a similar methodology to *Pinnuck* with a few exceptions. First, we use stocks only in the S&P/ASX 300 every December month-end. We use the S&P/ASX 300 as the universe in recognition of the skewed market capitalization distribution to the largest stocks in the Australian market, as evident in Table 1. Stocks beyond the largest 300 stocks very rarely fall into the tradeable universe for fund managers. We also use a 4/3/2 sort instead (approximately 10–11 stocks in each portfolio) because using the 5/4/3 portfolio sort Pinnuck (2003) uses will result in characteristic benchmarks with five stocks or less. Portfolios are value-weighted using index-weights (although similar results are found when value-weighting by market capitalization) and every month, the benchmark portfolios are rebalanced by each stock's month-end index weight and held for the next month. This is to avoid the characteristic benchmarks deviating from actual market weights (and, hence, from the S&P/ASX 300 return).

The *Broad* benchmark attempts to address the issue of too few stocks in each characteristic portfolio in *Index*, which might result in a high level of idiosyncratic risk in each portfolio. As such, it uses the same methodology as *Index* but uses broader benchmarks through using a 1/3/3 portfolio sort procedure (approximately 30 stocks in each portfolio). In this case, the sort on market capitalization is removed as the S&P/ASX 300 essentially consists of the largest stocks on the ASX.

The *Overlap* benchmark uses an overlapping portfolio methodology similar to Jegadeesh and Titman (1993). A stock enters a characteristic benchmark portfolio in a given month t if it meets the following data criteria: market capitalization and share price data for month $t-1$, book value data in the previous year or if the stock's current year reporting date is 3 or more months earlier than month $t-1$, the current year's book value, past 12 month returns and has a weight in the S&P/ASX 300 index for month $t-1$.[6]

The characteristic portfolios are formed as follows. At the end of each month (rather than just at every December month-end), all stocks that meet our data criteria are placed into portfolios using a 4/3/2 sorting procedure. Each portfolio is weighted using S&P/ASX 300 weights from the previous month-end

[6] Under ASX periodic disclosure rules for our sample period, an entity must disclose its accounts no later than 75 days after the end of its accounting period.

and held for 12 months, with monthly reweighting by month-end index weights. Therefore, in a given month, a stock's respective characteristic portfolio is the equal-weighted return of 12 overlapping characteristic portfolios.

The use of an overlapping benchmark, in contrast to the annually revised benchmark of Daniel *et al*. (1997) and Pinnuck (2003), allows for the incorporation of timely information into our benchmarks. In the Daniel *et al*. (1997) framework, a stock's style characteristics might be up to 12 months old. Therefore, a winner momentum stock 12 months ago might be a neutral momentum stock 6 months later. However, in our overlapping methodology, the latest characteristic information is used to form more timely benchmarks. To reduce noise in benchmarks from solely weighting on the past month's characteristic information, the past L month average benchmark of a stock is used. Therefore, if a stock is in transition from growth to value during the period, it will be considered on average a growth neutral stock.

A more practical reason for overlapping, and consequently monthly ranking, is to increase the sample population of stocks benchmarked. In a market benchmark such as the S&P/ASX 300 with a changing stock composition over time, stocks frequently enter and exit the benchmark intrayear.[7] As such, if we rank once yearly we might bias our benchmarks by only assessing surviving stocks which tend to be the largest stocks.

3.2. Calculation of characteristic-based benchmark measures

Following Daniel *et al*. (1997), characteristic selectivity (CS) is measured as the fund's gross return of the portfolio less the fund's value-weighted characteristic benchmark return as a result of the characteristics of stock holdings.[8] Mathematically, the monthly CS return for a fund over time period t is:

$$CS_t = \sum_{i=1}^{N} w_{i,t-1}(R_{i,t} - R_t^{bi,t-1}), \tag{1}$$

where $w_{i,t-1}$ is the weight of stock i in month $t-1$; $R_{i,t}$ is the monthly return of stock i in month t; and $R_t^{bi,t-1}$ is the monthly return of the matching characteristic benchmark portfolio to stock i at month $t-1$ in month t.

---

[7] In unreported results, a monthly updating benchmark in our sample period captures approximately 91 per cent of stocks on the ASX 300 by stock count and 92 per cent by market capitalization throughout the year. However, a December month-end annual ranking benchmark assessed in next year's November month-end (i.e. the last month-end before reranking occurs) on average captures only 76 per cent of stocks on the ASX 300 and 86 per cent by market capitalization.

[8] For funds holding option contracts, we follow Pinnuck (2003) and calculate the instantaneous equivalent underlying ordinary share position.

An important property unique to our measure is that by definition, holding the index portfolio will yield a zero CS measure because of the benchmark portfolio formation methodology. Therefore, a fund's holding is simultaneously being assessed against deviation from the S&P/ASX 300 index as well as against the characteristics of stocks.

By construct, the Daniel *et al*. (1997) components are a decomposition of a portfolio's raw return. One limitation of this decomposition is the requirement of a fund's past year holdings history in the characteristic timing (CT) and average style (AS) measures, which imposes data restrictions to our relatively short holdings history. To reduce this requirement, we merge the CT and AS measures to form the style return (SR) measure as:

$$SR_t = \sum_{i=1}^{N} w_{i,t-1} R_t^{bi,t-1}, \tag{2}$$

where the notations are the same as those used in equation (1). By definition, if all characteristic benchmark stocks are held using index weights (reweighted over the sum of all stock index weights in the characteristic benchmark), the SR measure equates to the implied market (IM) return, which is the return inferred by the characteristic benchmark:

$$IM_t = \sum_{n=1}^{N} w_{m,i,t-1} R_t^{bi,t-1}, \tag{3}$$

where $w_{m,i,t-1}$ is the 1 month lagged index weight in stock i.

Therefore, we can measure the SR of a fund in excess of the market, excess style (ES), as:

$$ES_t = SR_t - IM_t. \tag{4}$$

Excess style represents a concise measure of whether a fund is able to time or pick styles (or a mixture of both) over the market return.

In summary, our characteristic-based benchmark decomposes raw holding returns into IM, CS and ES returns:

$$R_{p,t} = CS_{p,t} + ES_{p,t} + IM_t. \tag{5}$$

3.3. Benchmark statistical measures on PAD funds and passive portfolio simulation

To compare how well each characteristic-based benchmark captures passive style, we use several statistical measures using two holding datasets: actual active equity fund holdings from the PAD and simulated passive portfolios following Kothari and Warner (2001).

In our first test using PAD data, we adopt two measures: tracking error and IM correlation. Tracking error is the annualized standard deviation of CS. Chan *et al*. (2006) assert that tracking error should be low if a benchmark portfolio aligns with the investment manager's investment mandate. To compare whether tracking error of our alternative benchmarks is statistically different to the *Pinnuck* benchmark, we use the Levene's test for homogeneity in variances. In addition, we test for statistical significance in differences of CS between the standard *Pinnuck* benchmark and our alternatives using Newey–West *t*-statistics.[9]

Implied market correlation is measured as the correlation of the monthly IM (i.e. the IM return from equation 3) with the actual return of the S&P/ASX 300A to measure the deviation of the characteristic benchmark. Ideally, correlation of IM to the S&P/ASX 300A index should be as close to 100 per cent as possible.

Simulated passive portfolios to test benchmarks are formed following Kothari and Warner (2001). As these portfolios are by design passive (i.e. have an *ex ante* zero alpha), this test measures the benchmark's ability to correctly infer zero alpha to these simulated portfolios. Every month, stocks in the S&P/ASX 300 that satisfy the *Index* benchmark criteria are independently sorted into two groups by market capitalization, book-to-market and prior 1 year return to form six groups. These portfolios simulate fund manager investment styles: small cap, large cap, growth, value, momentum and contrarian funds. In each group, 50 stocks are randomly selected by equal probability, or based on market capitalization, to form a portfolio.[10] The portfolios are held equal-weight or value-weight, and not monthly rebalanced (i.e. buy and hold) for 12 months. At the end of months 12 and 24, the portfolios are reformed to generate a time series of returns for 36 months for a given passive portfolio. This results in 24 unique passive investment style combinations (six styles, two selection methods and two weighting methods). We form portfolios with holdings months matching the PAD sample period from January 1995 to June 1999 month-ends (the last holding month of a passive portfolio formed at the start of June 1999 being June 2002) to form 54 passive portfolios of 36 months in length for each style combination. As an out-of-sample test of the benchmarks, we also form portfolios at month-ends from July 1999 to November 2003 (the last portfolio return month being December 2006, and the end of our SPPR dataset) for a total of 53 portfolios per style combination.

In addition to the four characteristic-based benchmarks, we also assess the regression-based Carhart model using the SPPR universe of stocks and only the S&P/ASX 300 stocks. We follow the methodology of Fama and French (1993) and Carhart (1997) to form the factor loadings, and for brevity do not detail the

[9] In all our tests using Newey–West *t*-statistics, we use lags equalling $n^{0.25}$, where n is equal to the number of months a measure is calculated over.

[10] In our PAD sample, the median fund holds 48 stocks on average of its sample period.

methodology. We use the S&P/ASX 300 accumulation index as our market proxy and the monthly return of the 13 week treasury note from the CRIF SPPR database as our risk-free rate.

We use several statistical measures to compare the benchmarks. First, we measure the frequency rate at which the null hypothesis of zero alpha is rejected at the 5 per cent level, using a two-tailed test, for the passive portfolios in a given style combination. Ideally, the rejection frequency is zero for a benchmark.

Similar to our tests using the PAD data, the average tracking error is measured for each of the portfolios in each style combination. Following Chan *et al.* (2006), we measure the tracking error of the Carhart model as the standard deviation of a month's actual return less the model's expected return estimated without that month's observation (to prevent overfitting of the model). Following Kothari and Warner (2001), we also measure the Newey–West standard error of mean alpha across the passive portfolios in each style combination. For all measures, we calculate the average difference across style combinations of the *Pinnuck* benchmark with the alternative benchmarks.

4. Results

4.1. Unadjusted returns

To highlight the importance of using similar frequency data to reduce standard errors, in Table 2 we calculate an 'implied' S&P/ASX 300 accumulation index return in accordance with equation (3), and compare it to the actual index return (using month-end price levels). Panel A of Table 2 reports the annualized average monthly returns of the S&P/ASX 300 accumulation index from index levels (ASX 300A) and from S&P/ASX 300 market benchmark weights (Implied ASX 300A). We also measure the value-weighted PAD portfolio return. The returns of the Implied ASX 300A and value-weighted PAD fund are calculated by using month-end weights at $t - 1$ and holding for month t.

During this period, the Implied ASX 300A return of 11.41 per cent per year is approximately equal to the ASX 300A return. Therefore, intramonth fluctuations in market weights do not appear to greatly affect the return of the market.[11]

Our calculation of the excess PAD return of PAD less Implied ASX 300A and PAD less ASX 300A return is more revealing. Despite the economically significant magnitude of approximately 3 per cent per year, the statistical significance greatly differs. The PAD less Implied ASX 300A has a t-statistic of 2.90, higher than that of PAD less ASX 300A of 2.24. This difference can be seen in the Pearson correlation matrix of monthly returns in Panel B. There is a 98.06 per cent correlation between Implied ASX 300A and PAD, but the correlation between Actual Market and PAD is only 96.13 per cent. Therefore,

[11] One additional discrepancy is that we do not use the returns of non-ordinary stocks as this is unavailable in the CRIF SPPR.

Table 2
Annualized monthly average returns of holding returns

Panel A: Raw return averages

Actual ASX 300A return	Implied ASX 300A market	Implied PAD value-weighted holdings	PAD less Actual ASX 300A	PAD less Implied ASX 300A
11.41***	11.41***	14.40***	3.00**	2.99**
(3.32)	(3.33)	(4.03)	(2.24)	(2.90)

Panel B: Pearson correlation matrix of returns

	Actual ASX 300A	Implied ASX 300A
Implied ASX 300A	0.9900	
PAD funds	0.9613	0.9806

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. Panel A presents the raw annualized monthly average market and Portfolio Analytics Database (PAD) returns from January 1995 to June 2002. The return of the Australian Stock Exchange (ASX) 300 Accumulation Index is calculated using month-end price levels. Standard & Poor's (S&P)/ASX 300 Accumulation Index Implied Return is calculated using lagged month index weights multiplied by the current month return. Implied PAD Return holdings is calculated using lagged month weights of value-weighted stock holdings of all PAD managers multiplied by the current month's return. Panel B shows the Pearson correlation matrix of returns. Newey–West *t*-statistics are in parentheses.

it is of importance to use the IM return when calculating our ES measure. The correlation between the ASX 300A and Implied ASX 300 is 99.00 per cent, suggesting the implied return accurately describes the returns of the actual ASX despite the slight discrepancies. The importance of this is shown in later sections when we test the correlation of the Implied ASX 300A return from characteristic benchmark weights against the actual ASX 300A return.

4.2. Characteristic-based benchmarks on PAD funds

This section compares the standard characteristic benchmark, *Pinnuck*, to the *Index*, *Broad* and *Overlap* benchmarks described in Section 3.1. Table 3 reports the results of our decomposition of PAD fund holding returns into CS, ES, unadjusted return (Raw), IM, correlation of IM to the S&P/ASX 300 (Corr.) and tracking error (TE) using the different methodologies. All measures (except Corr.) are in per cent per year. We also report the CS difference (ΔCS) of an alternative benchmark to the *Pinnuck* benchmark (ΔCS) and critical *p*-value of the Levene's test for homogeneity of variances of a benchmark and *Pinnuck* (Lev.) are also reported.

For our initial test in Panel A, we adopt the standard characteristic benchmark methodology following Pinnuck (2003). We find a CS of 1.87 per cent per

Table 3
Characteristic-based benchmark performance measures

Panel A: Pinnuck (2003) benchmark

CS	ES	SR	Raw	IM	Corr.	TE
1.87***	2.36**	12.26***	14.14***	9.91**	0.9434	2.19
(2.73)	(2.54)	(3.31)	(3.81)	(2.55)		

Panel B: S&P/ASX 300 4/3/2 Portfolio Sorts (Index)

CS	ES	SR	Raw	IM	Corr.	TE	ΔCS	Lev.
1.08**	1.37**	13.28***	14.36***	11.91***	0.9836	1.43	–0.79	0.0024
(2.12)	(2.52)	(3.63)	(3.83)	(3.30)			(–1.25)	

Panel C: S&P/ASX 300 1/3/3 Portfolio Sorts (Broad)

CS	ES	SR	Raw	IM	Corr.	TE	ΔCS	Lev.
1.17**	1.28***	13.19***	14.36***	11.91***	0.9836	1.48	–0.70	0.0012
(2.11)	(2.97)	(3.61)	(3.83)	(3.30)			(–1.10)	

Panel D: S&P/ASX 300 Overlapping Benchmark 4/3/2 Portfolio Sorts (Overlap)

CS	ES	SR	Raw	IM	Corr.	TE	ΔCS	Lev.
1.79***	0.95**	12.62***	14.41***	11.67***	0.9845	1.34	–0.08	0.0002
(3.33)	(2.30)	(3.46)	(3.83)	(3.21)			(–0.13)	

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. This table reports the time series average monthly annualized characteristic selectivity (CS), excess style (ES), style return (SR), raw return (Raw), implied market (IM), tracking error (TE), difference of CS to the *Pinnuck* benchmark (ΔCS) and critical *p*-value of the Levene's test for homogeneity of variances of the benchmark and *Pinnuck* (Lev.) for value-weighted Portfolio Analytics Database (PAD) funds from January 1995 to June 2002 using different characteristic benchmark methodologies. Corr. is the correlation of IM to the return of the Standard & Poor's/Australian Stock Exchange (S&P/ASX) 300 Accumulation Index from price levels. Newey–West *t*-statistics are in parentheses.

year, statistically significant at the 5 per cent level. This is slightly lower than the sample used by Pinnuck (2003) of approximately 2 per cent a year, although he uses a different sample of funds and a sample period from June 1990 to June 1997. Note that the 9.91 per cent per year IM is also 1.42 per cent lower than that of the S&P/ASX 300A of 11.41 per cent reported in Panel A of Table 2 (as a result of using the entire ASX sample to form characteristic benchmarks rather than the investable benchmark S&P/ASX 300). The reported value-weighted return of all stocks in the CRIF SPPR during this period is 10.05 per cent per year (t = 2.87) confirming that the S&P/ASX 300A outperformed the broader benchmark

during this period.[12] As a result, the correlation of IM to the S&P/ASX 300A is only 94.34 per cent, lower than the 99.00 per cent reported in Panel B of Table 2.

In Panel B of Table 3, using the *Index* benchmark yields CS of 1.08 per cent per year ($t = 2.12$), which is 0.79 per cent lower than that reported of the *Pinnuck* benchmark, although this difference is not statistically different. The IM correlation of 98.36 per cent is also higher and tracking error significantly reduced to 1.44 compared with the *Pinnuck* benchmark. The difference in tracking error to *Pinnuck* is statistically significant as shown in the *p*-value of the Levene's test of 0.24 per cent.

Panel C of Table 3 reports results using the *Broad* benchmark where the same methodology as *Index* is used except for removing the portfolio sort on size. The statistically significant CS of 1.17 per cent and tracking error of 1.57 per cent are both higher than for the *Index* benchmark. In unreported results, the CS and tracking error, however, are not statistically different to the *Index* benchmark.

Panel D of Table 3 reports results using the overlapping methodology as described in Section 3.1, which uses up-to-date characteristic information and is able to benchmark stocks that enter a portfolio in the middle of the year. The benchmark's correlation to the market of 98.45 per cent is slightly higher than that of the *Index* and *Broad* benchmarks and has lower tracking error of 1.34 per cent and higher CS of 1.79 per cent. The lower tracking error is statistically different to the *Pinnuck* benchmark, although not different to the *Index* and *Broad* benchmarks. The higher CS, although not statistically different to the *Pinnuck* measure, is statistically different to *Broad* and *Overlap* (unreported). To improve comparability of the *Overlap* benchmark to *Index* and *Broad*, we use only stocks in the *Index* benchmark to remove stocks entering portfolios in the middle of the year ('mid-entry' stocks) from the *Overlap* benchmark. In unreported results, we find a CS without mid-entry stocks of 1.72 per cent ($t = 3.02$), which is not statistically different to the *Overlap* benchmark and is still statistically significant with respect to the *Index* and *Broad* CS measures. This suggests that the difference in benchmark measures is not due to different stocks being assessed.

The discrepancies in CS and tracking error might be due to a benchmark not being able to adequately capture the characteristic returns of a fund's style. In Table 4, we repeat the same analysis except sort the PAD funds by self-reported style to see the adequacy of the benchmarks in capturing by fund style. The CS measures using the various benchmarks reported in Panel A show disagreement in the statistical significance and magnitude of CS within a fund style. For example, for Value funds, although all benchmarks show statistically significant CS, *Index* reports a measure of 2.04 per cent per year, and *Pinnuck* (2.89 per cent), *Broad* (2.51 per cent) and *Overlap* (2.76 per cent) vary. As approximately

[12] Again, the discrepancy between our reported 9.44 per cent market return with that of the CRIF SPPR is due to filtering for stocks that meet our data requirements.

Table 4
Comparison of benchmark measures of CS and tracking error by fund style

Panel A: CS and CS difference measures

Style	*Pinnuck*	*Index*	*Broad*	*Overlap*	*Index – Pinnuck*	*Broad – Pinnuck*	*Overlap – Pinnuck*
GARP	0.52	–0.08	–0.26	0.72	–0.60	–0.78	0.20
Growth	3.12**	1.99*	1.74	2.56**	–1.12	–1.38	–0.56
Other	1.27	0.37	0.57	0.57	–0.90	–0.70	–0.71
Style neutral	1.89*	2.27***	2.46**	2.03***	0.37	0.57	0.14
Value	2.89***	2.04**	2.51***	2.76***	–0.85	–0.38	–0.13

Panel B: Tracking error (per cent per year) and Levene's test critical p-*values*

Style	*Pinnuck*	*Index*	*Broad*	*Overlap*	*Index/ Pinnuck*, *p*-value	*Broad/ Pinnuck*, *p*-value	*Overlap/ Pinnuck*, *p*-value
GARP	2.38	1.74	1.73	1.72	0.007	0.005	0.005
Growth	3.71	2.83	2.68	2.57	0.015	0.005	0.003
Other	2.81	1.62	1.80	1.56	0.001	0.003	0.001
Style Neutral	3.53	3.16	2.98	2.64	0.552	0.321	0.106
Value	2.53	2.41	2.31	2.27	0.692	0.457	0.352

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. This table reports the characteristic selectivity (CS) and tracking error using alternative characteristic benchmarking methodologies on the value-weighted holdings of Portfolio Analytics Database (PAD) funds by self-reported style from January 1995 to June 2002. Panel A reports the average annualized monthly CS and CS differences using Newey–West *t*-statistics of the *Pinnuck, Index, Broad* and *Overlap* characteristic benchmarks detailed in Section 3.1. Panel B reports the annualized tracking error and Levene's test for homogeneity of variances critical *p*-values between the *Index, Broad* or *Overlap* benchmark against the *Pinnuck* benchmark.

40 per cent of aggregate PAD holdings are in Value funds, this partially explains why the aggregate CS using *Index* of 1.08 per cent reported in Table 2 is lower than that of the *Pinnuck* (1.87 per cent) and *Overlap* (1.79 per cent) benchmarks. Similarly for Growth, CS measures range from 3.12 per cent, which is statistically significant using *Pinnuck*, to 1.74 per cent, which is not significant using *Broad*. However, the statistical difference between an alternative benchmark's CS to the *Pinnuck* benchmark (*Index – Pinnuck*, *Broad – Pinnuck* and *Overlap – Pinnuck*) and other benchmarks (unreported) is not statistically significant except for differences in CS for *Index* and *Overlap*, *Broad* and *Overlap* for GARP funds. As approximately 35 per cent of total PAD funds is in GARP, this suggests that the differences in CS of *Overlap* to *Index* and *Broad* is due to *Overlap* assigning a higher although not statically significant CS to GARP funds. Therefore, there is no clear upward or downward bias in CS of the benchmarks despite magnitude differences.

For tracking error as reported in Panel B, *Overlap* has the lowest measure across all fund styles compared with all other benchmarks. However, the difference is only significant compared to the *Pinnuck* benchmark. The Levene's test *p*-values show that tracking error differences are statistically significant for *Index*/*Pinnuck*, *Broad*/*Pinnuck* and *Overlap*/*Pinnuck* pairs for GARP, Growth and Other funds. However, in all other styles and unreported pairings of *Index*, *Broad* or *Overlap*, the differences are not statistically significant. This suggests that the lower tracking errors of *Index*, *Broad* and *Overlap*, while improving on the *Pinnuck* methodology, are indistinguishable in superiority within the alterative benchmarks. However, the problem remains of the varying measures of CS in aggregate and across fund styles, and which of the alternative benchmarks is the 'correct' measure in terms of magnitude and statistical significance. In the next section, we turn to using simulated passive portfolios to test the validity of the measures.

4.3. Passive portfolio simulation

This section tests the characteristic benchmarks using passive-style simulated portfolios. Our previous tests using PAD have inherent difficulties in inference testing, as the abnormal return *ex ante* is unknown and is sample and time specific. Table 5 reports our results using the four characteristic benchmarks and two variants of the Carhart model, C4 and C4 ASX 300, on the 24 style combinations. Average alpha (Panel A), percentage of portfolio rejecting the null of zero alpha at the 5 per cent level (Panel B), average tracking error (Panel C) and average Newey–West standard errors (Panel D) across the 54 portfolios in each style are reported.[13] In Panels A and B, we find that all benchmarks do not assign a near zero alpha to the passive-style portfolios, and the *Pinnuck* benchmark has the lowest rejection rate of the null hypothesis of zero alpha. Although the mean alpha varies greatly in a particular style combination, the cross-sectional average for all benchmarks is not statistically significant with the exception for *Broad* being –1.17 per cent per year and statistically significant and, hence, suggesting some downward bias in the alpha measure. In addition, the averages of *Index* (–0.30 per cent) and *Overlap* (–0.17 per cent) are closer to zero and statistically different to *Pinnuck*. This suggests that the *Index* and *Overlap* benchmarks overall are the most alpha neutral, compared with the other benchmarks. In rejection rates, the *Pinnuck* benchmark has the lowest rate at 6.56 per cent of portfolios, with *Index* and *Broad* benchmark's rejection rates not statistically significant, and the *Overlap* and Carhart models having statistically significant and higher rejection rates. Tracking error and Newey–West

[13] For conciseness, only cross-sectional averages and cross-sectional differences of the measures across style combinations are reported. Average measures of portfolios in each style combination are available on request.

Table 5
Comparison of benchmarks using simulated passive style portfolios from January 1995 to June 1999

Statistic	*Pinnuck*	*Index*	*Broad*	*Overlap*	C4	C4 ASX 300
Mean alpha	0.42	–0.30	–1.17**	–0.17	0.41	0.30
	(1.18)	(–1.30)	(–2.41)	(–0.70)	(0.64)	(0.45)
ΔPinnuck		–0.72***	–1.60***	–0.59***	–0.01	–0.12
		(–5.95)	(–13.02)	(–3.77)	(–0.07)	(–0.61)
Rejection rate	6.56	8.33	7.02	11.34	11.73	11.42
ΔPinnuck		1.77	0.46	4.78***	5.17***	4.86***
		(1.25)	(0.33)	(4.84)	(3.57)	(3.83)
Tracking error	5.43	4.56	5.66	3.88	7.63	7.75
ΔPinnuck		–0.87***	0.23	–1.55***	2.20***	2.32***
		(–12.39)	(0.75)	(–24.38)	(11.09)	(11.51)
NW standard error	0.15	0.12	0.14	0.11	0.23	0.23
ΔPinnuck		–0.03***	–0.01	–0.04***	0.08***	0.09***
		(–5.18)	(–1.11)	(–10.85)	(13.42)	(13.44)

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. At the end of every month from January 1995 to June 1999 (54 months), stocks in the Standard & Poor's/Australian Stock Exchange (S&P/ASX) 300 that satisfy the *Index* benchmark criteria are ranked and independently sorted into two groups by market capitalization, book-to-market and prior 1 year return to form six groups. These portfolios simulate fund manager investment styles: small cap (Small), large cap (Large), growth (Growth), value (Value), momentum (Momentum) and contrarian (Contrarian) funds. In each group, 50 stocks are randomly selected by equal probability (Choice = Equal) or based on market capitalization (Choice = Cap) to form a portfolio. The portfolios are held equally (Weight = EW) or value-weighted (Weight = VW) and not rebalanced (i.e. buy and hold) for 12 months. At the end of months 12 and 24, the portfolios are reformed to form a time series of returns for 36 months for a given passive portfolio. This results in 24 unique passive investment style combinations (six styles, two selection methods and two weighting methods). The portfolios are assessed against the four characteristic-based benchmarks detailed in Section 3.1 and two variants of the Carhart model, one using stocks in the CRIF SPPR universe (C4) and the other using only S&P/ASX 300 stocks (C4 ASX 300). For each portfolio, we calculate the time series mean monthly alpha, whether this alpha rejects the null hypothesis of zero alpha at the 5 per cent level (using a two-tailed test), tracking error and Newey–West standard error of the alpha. Using these measures, we then calculate for the 54 portfolios in each style combination, the average mean alpha in per cent per year (Mean alpha), percentage portfolios rejecting the null hypothesis of zero alpha at the 5 per cent level, using a two-tailed test (Rejection rate), Tracking error in per cent per year and Newey–West standard error in per cent per year (NW standard error). We also measure the cross-sectional average measure across style combinations (Average) and the difference of a benchmark's cross-sectional average to the *Pinnuck* cross-sectional average (ΔPinnuck). Newey–West *t*-statistics are in parentheses.

standard errors are lower for *Index* and *Overlap* measures compared with the *Pinnuck* benchmark. In Panel C, we find that differences in tracking error to *Pinnuck* of *Index* and *Overlap* of –0.87 and –1.55 per cent, respectively, are statistically significant. This is consistent with our findings using PAD in the above sections. Similarly, the Newey–West standard errors we report in Panel D are lower. Interestingly, the tracking error and Newey–West standard errors of the Carhart models, C4 and C4 ASX 300, are higher and statistically different

Table 6
Out-of-sample testing of benchmarks using simulated passive portfolios from July 1999 to November 2003

Statistic	*Pinnuck*	*Index*	*Broad*	*Overlap*	C4	C4 ASX 300
Mean alpha	1.06**	0.41	1.10**	0.34	2.50***	2.34***
	(2.67)	(1.29)	(2.69)	(1.07)	(4.39)	(4.20)
ΔPinnuck		–0.65***	0.03	–0.72***	1.44***	1.28***
		(–7.16)	(0.42)	(–7.37)	(13.31)	(11.93)
Rejection rate	8.49	8.65	8.49	10.06	15.09	14.07
ΔPinnuck		0.16	0.00	1.57	6.60***	5.58***
		(0.13)	(0.00)	(1.00)	(3.48)	(3.04)
Tracking error	4.73	3.90	5.18	3.52	6.47	6.50
ΔPinnuck		–0.83***	0.44***	–1.21***	1.73***	1.77***
		(–7.40)	(3.02)	(–8.95)	(12.99)	(12.94)
NW standard error	0.11	0.09	0.14	0.09	0.21	0.21
ΔPinnuck		–0.02***	0.03***	–0.02***	0.10***	0.10***
		(–3.37)	(2.81)	(–3.91)	(11.72)	(11.91)

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. At the end of every month from July 1999 to November 2003 (53 months), stocks in the Standard & Poor's/Australian Stock Exchange (S&P/ASX) 300 that satisfy the *Index* benchmark criteria are ranked and independently sorted into two groups by market capitalization, book-to-market and prior 1 year return to form six groups. These portfolios simulate fund manager investment styles: Small cap, large cap, growth, value, momentum and contrarian funds. In each group, 50 stocks are randomly selected by equal probability or based on market capitalization to form a portfolio. The portfolios are held equally or value-weighted and not rebalanced (i.e. buy and hold) for 12 months. At the end of months 12 and 24, the portfolios are reformed to form a time series of returns for 36 months for a given passive portfolio. This results in 24 unique passive investment style combinations (six styles, two selection methods and two weighting methods) and 53 portfolios in each style combination. The portfolios are assessed against the four characteristic-based benchmarks detailed in Section 3.1 and two variants of the Carhart model, one using stocks in the CRIF SPPR universe (C4) and the other using only S&P/ASX 300 stocks (C4 ASX 300). For each portfolio, we calculate the time series mean monthly alpha, whether this alpha rejects the null hypothesis of zero alpha at the 5 per cent level (using a two-tailed test), tracking error and Newey–West standard error of the alpha. Using these measures, we then calculate for the 53 portfolios in each style combination, the average mean alpha in per cent per year (Mean alpha), percentage of portfolios rejecting the null hypothesis of zero alpha (Rejection rate), Tracking error in per cent per year and Newey–West standard error in per cent per year (NW standard error). This table reports cross-sectional averages of these measures across the 24 style combinations. We also measure the cross-sectional average measure across style combinations (Average) and the difference of a benchmark's cross-sectional average to the *Pinnuck* cross-sectional average (ΔPinnuck). Newey–West *t*-statistics are in parentheses.

to *Pinnuck* verifying the assertion of Daniel *et al*. (1997) that regression-based analysis has higher standard errors in the measurement of alpha.

As a further robustness test, we repeat the test out-of-sample for portfolios formed after the PAD period from July 1999 to November 2003. Table 6 reports our results. Again for conciseness, we only report cross-sectional averages and cross-sectional differences of the measures. We find that the results are

generally consistent to our previous findings, with the *Index* and *Overlap* benchmarks having CS closest to zero and lower tracking error and Newey–West standard errors, and all these measures being statistically different to *Pinnuck*. In addition, we find that although the *Overlap* rejection rate remains higher than that of *Pinnuck*, it is not statistically significant. In addition, we find that the rejection rates for the Carhart models are statistically different and higher compared with *Pinnuck*, consistent with our above findings and in the literature (e.g. Kothari and Warner, 2001; Chan *et al*., 2006) of higher error in regression-based models.

Finally, we compare differences in measures between the alternative benchmarks *Index*, *Broad* and *Overlap* for the two sample periods in Table 7. We find that the *Index* and *Overlap* benchmarks (*Index – Overlap*) are not statistically different for mean alpha and Newey–West standard errors. The rejection rate is statistically different and lower for the *Index* benchmark only in the first period, whereas tracking error is higher and statistically different for both periods compared with the *Overlap* benchmark. In comparison to the *Broad* benchmark (*Index – Broad* and *Overlap – Broad*), we find that *Broad* has statistically different alpha and higher and statistically different tracking and Newey–West standard error, although *Broad* has a statistically different and lower rejection rate than the *Pinnuck* benchmark in the first period. Taken together, this suggests the *Broad* benchmark has lower statistical power compared with the *Index* and *Overlap* benchmarks.

5. Conclusion

We explore the application of characteristic benchmarks and propose modifications to the standard characteristic benchmark methodology. The methodology we propose and contribute to the literature better enables a more precise measurement of stock selection ability through the capture of characteristic stock returns. In forming this benchmark, we consider issues that: (i) incorporate more timely characteristic information in the formation of the characteristic portfolios; (ii) matching characteristic portfolios to migrating stocks; (iii) improves a performance analyst's ability to benchmark stocks entering the market index intrayear; and (iv) assigning zero alpha to a market index replicating strategy (such as the S&P/ASX 300 Index).

Applying this modified benchmark to active Australian fund manager monthly holdings, we find a near halving in tracking error volatility of the overlapping benchmarks (i.e. more frequent updating of characteristic information in benchmarks) and also lower tracking error when benchmarking by fund style and stock characteristics compared with the standard characteristic benchmark following Pinnuck (2003).

Our results also contribute to the performance evaluation literature when testing the benchmark's ability against simulated passive-style portfolios mimicking the investment styles of fund managers. Statistically different and

Table 7
Differences of alternative benchmark statistical measures

Statistic	Period	*Index – Overlap*	*Index – Broad*	*Overlap – Broad*
Mean alpha (% per year)	First	–0.13	0.87***	1.00***
		(–0.92)	(4.21)	(3.78)
	Second	0.07	–0.69***	–0.75***
		(0.86)	(–5.03)	(–5.79)
Rejection rate (%)	First	–3.01**	1.31	4.32***
		(–2.19)	(0.83)	(3.61)
	Second	–1.41	0.16	1.57
		(–1.11)	(0.20)	(1.11)
Tracking error (% per year)	First	0.68***	–1.10***	–1.77***
		(7.78)	(–4.25)	(–5.61)
	Second	0.37***	–1.28***	–1.65***
		(8.35)	(–5.79)	(–6.55)
Newey–West standard error (% per year)	First	0.01*	–0.02***	–0.03***
		(1.87)	(–6.00)	(–4.75)
	Second	0.00	–0.05***	–0.05***
		(0.31)	(–3.55)	(–4.60)

***, ** and * denote statistical significance at the 1, 5 and 10 per cent levels, respectively. At the end of every month from January 1995 to June 1999 (54 months, first period) and July 1999 to November 2003 (53 months, second period), stocks in the Standard & Poor's/Australian Stock Exchange 300 that satisfy the *Index* benchmark criteria are ranked and independently sorted into two groups by market capitalization, book-to-market and prior 1 year return to form six groups. These portfolios simulate fund manager investment styles: Small cap, large cap, growth, value, momentum and contrarian funds. In each group, 50 stocks are randomly selected by equal probability or based on market capitalization to form a portfolio. The portfolios are held equally or value-weighted and not rebalanced (i.e. buy and hold) for 12 months. This results in 24 unique passive investment style combinations (six styles, two selection methods and two weighting methods). At the end of months 12 and 24, the portfolios are reformed to form a time series of returns for 36 months for a given passive portfolio. The portfolios are assessed against the four characteristic-based benchmarks detailed in Section 3.1. For each portfolio, we calculate the time series mean monthly alpha, whether this alpha rejects the null hypothesis of zero alpha at the 5 per cent level (using a two-tailed test), tracking error and Newey–West standard error of the alpha. Using these measures, we then calculate for the 53 portfolios in each style combination, the average mean alpha in per cent per year (Mean alpha), percentage of portfolios rejecting the null hypothesis of zero alpha (Rejection rate), tracking error in per cent per year (Tracking error) and Newey–West standard error in per cent per year (NW standard error). We then calculate the cross-sectional averages of these measures across the 24 style combinations. The table reports the average cross-sectional differences of the measures between the *Index*, *Broad* and *Overlap* benchmarks in the two periods. Newey–West *t*-statistics are in parentheses.

lower tracking error and Newey–West standard errors, and also an average alpha closer to zero, are achieved compared to using the standard benchmark. The same improvements are found compared with using a market neutral benchmark that does not control for size (to ensure that more stocks and less idiosyncratic risk are in each characteristic portfolio), although only improved tracking error is achieved in comparison to a benchmark that is only market neutral. We also

verify that the characteristic benchmark methodology has superior statistical properties compared to the regression-based Carhart model.

Our findings show that simple modifications in the characteristic benchmark methodology improves the ability of the benchmark to better capture characteristic stock returns and, therefore, more accurately measure stock selection ability. More specifically, focused benchmarks within the fund manager's investable domain provide improved quantification of genuine managerial ability and stock selection skill. However, an important caveat remains: in our tests of simulated passive portfolios, we find that the standard and modified characteristic benchmarks reject the null hypothesis of zero alpha, on average, approximately 8–10 per cent of the time, suggesting that the benchmarks still remain less than perfect in stock selection detection. Nonetheless, our findings have important implications for future research in considering the choice of benchmarking methodology by which active investment managers are scrutinized.

References

Avramov, D., and R. Wermers, 2006, Investing in mutual funds when returns are predictable, *Journal of Financial Economics* 81, 339–377.

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

Chan, L. K. C., S. G. Dimmock, and J. Lakonishok, 2006, Benchmarking money manager performance: Issues and evidence, working paper (University of Illinois at Urbana-Champaign, Urbana, IL).

Chen, Z., and P. J. Knez, 1996, Portfolio performance measurement: theory and applications, *Review of Financial Studies* 9, 511–555.

Coval, J. D., and T. J. Moskowitz, 2001, The geography of investment: informed trading and asset prices, *Journal of Political Economy* 109, 811–841.

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

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *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, 55–84.

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

Gallagher, D. R., and A. Looi, 2006, Trading behaviour and the performance of daily institutional trades, *Accounting and Finance* 46, 125–147.

Green, R. C., 1986, Benchmark portfolio inefficiency and deviations from the security market line, *Journal of Finance* 41, 295–312.

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

Gruber, M. J., 1996, Another puzzle: the growth in actively managed mutual funds, *Journal of Finance* 51, 783–810.

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

Jegadeesh, N., and S. Titman, 2001, Profitability of momentum strategies: an evaluation of alternative explanations, *Journal of Finance* 56, 699–720.

Jensen, M. C., 1968, The performance of mutual funds in the period 1945–64, *Journal of Finance* 23, 389–416.

Kacperczyk, M., C. Sialm, and L. U. Zheng, 2005, On the industry concentration of actively managed equity mutual funds, *Journal of Finance* 60, 1983–2011.

Kothari, S. P., and J. B. Warner, 2001, Evaluating mutual fund performance, *Journal of Finance* 56, 1985–2010.

Lehmann, B. N., and D. M. Modest, 1987, Mutual fund performance evaluation: a comparison of benchmarks and benchmark comparisons, *Journal of Finance* 42, 233–265.

Malkiel, B. G., 1995, Returns from investing in equity mutual funds 1971–91, *Journal of Finance* 50, 549–572.

Pástor, L., and R. F. Stambaugh, 2002a, Investing in equity mutual funds, *Journal of Financial Economics* 63, 351–380.

Pástor, L., and R. F. Stambaugh, 2002b, Mutual fund performance and seemingly unrelated assets, *Journal of Financial Economics* 63, 315–349.

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

Roll, R., 1977, A critique of the asset pricing theory's tests part I: on past and potential testability of the theory, *Journal of Financial Economics* 4, 129–176.

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

Wermers, R., 2000, Mutual fund performance: an empirical decomposition into stock-picking talent, style, transactions costs, and expenses, *Journal of Finance* 55, 1655–1703.