Critical Finance Review, 2012, 2: 1–48

Should Benchmark Indices Have Alpha? Revisiting Performance Evaluation*

Martijn Cremers[1], Antti Petajisto[2], and Eric Zitzewitz[3]

[1]*University of Notre Dame, Mendoza College of Business, Notre Dame, IN 46556; mcremers@nd.edu*
[2]*Corresponding author. NYU Stern, 44 W 4th St, Suite 9-190, New York, NY 10012-1126; antti.petajisto@stern.nyu.edu*
[3]*Dartmouth College and NBER, 6016 Rockefeller Hall, Hanover, NH 03755; ericz@dartmouth.edu*

ABSTRACT

Standard Fama-French and Carhart models produce economically and statistically significant nonzero alphas, even for passive benchmark indices such as the S&P 500 and Russell 2000. We find that these alphas primarily arise from the disproportionate weight that the Fama-French factors place on small value stocks, which have performed well, and from the CRSP value-weighted market index, which is historically a downward-biased benchmark for U.S. stocks due to the inclusion of other types of securities such as closed-end funds. We propose small methodological changes to the Fama-French factors to eliminate the nonzero alphas, and we also propose factor models based on common and tradable benchmark indices. Both kinds of alternative models improve performance evaluation of actively managed portfolios with the index-based models exhibiting the best performance.

* We are grateful for comments by Yakov Amihud, Nick Barberis, Lauren Cohen, Wayne Ferson, Ken French, John Griffin, Ilan Guedj, Ron Kaniel, Michael Lemmon, Jonathan Lewellen, Juhani Linnainmaa, Raj Mehra, Claudia Moise, Lubos Pastor, Christopher Polk, Tarun Ramadorai, Pedro Santa-Clara, Clemens Sialm, Laura Starks, Michael Stutzer, Paul Tetlock, Ivo Welch, two anonymous referees, as well as conference participants at AFA 2010 Annual Meeting, EFA 2009 Annual Meeting, FIRS 2009 Conference, FMA 2009 European Conference, London School of Economics, NBER Asset Pricing Meeting, Utah Winter Finance Conference, and seminar participants at AQR Capital, Arizona

ISSN 2164-5744; DOI 10.1561/104.00000007

1 Introduction

Practitioners typically evaluate money managers by comparing their returns to benchmark indices such as the S&P 500 for large-cap stocks and the Russell 2000 for small-cap stocks. In contrast, academics use the Carhart four-factor model and the Fama-French (FF) three-factor model as the standard benchmarks. Our paper provides evidence that the practitioner and academic approaches can yield very different results because the academic factor models assign large nonzero alphas for extended periods of time to passive benchmark indices.

For example, regressing the excess returns of the S&P 500 index (including dividends) on the Carhart four-factor model yields an annual alpha of 0.82% ($t = 2.78$) over our sample period from 1980 to 2005, which is a commonly studied time period due to the availability of mutual fund holdings data starting in 1980. The Russell 2000 annual alpha is -2.41% ($t = -3.21$). A portfolio that is long the S&P 500 Growth index and short the Russell 2000 Growth index has an impressive annual alpha of 5.23% ($t = 4.23$). Hence, even pure index funds that track common benchmark indices appear to have significant positive or negative "skill." Yet these benchmark indices represent broad, well-diversified, and passive portfolios that almost by definition should have zero abnormal returns or alphas — after all, the S&P 500 and Russell 2000 together cover about 85% of the U.S. equity market value and are the two most common benchmark indices for fund managers.

This problem is not limited to common benchmark indices or the time periods over which these indices have existed. When we examine the Carhart and Fama-French alphas of size and book-to-market (BM) decile portfolios, we find that the largest size decile portfolios have positive alphas, the small-cap portfolios tend to have negative alphas, and the differences are more extreme for growth stocks. As with the indices, we can reject the

State University, Dartmouth College, Harvard Business School, Helsinki School of Economics, HKUST, INSEAD, Numeric Investors, NYU Stern, UC Irvine, University of Alabama, University of Amsterdam, University of Colorado, University of Massachusetts at Amherst, University of Michigan, University of Texas at Austin, University of Toronto, Wharton, and Yale School of Management. We also thank Frank Russell Co, Standard and Poor's, Dow Jones Wilshire, and Morningstar for providing us with data. We gratefully acknowledge the financial support of the Q-Group, as well as the EFA Best Paper award of the Commonfund Institute, the FMA Europe Best Paper award, and the Roger F. Murray Prize of the Q-Group.

joint null hypothesis that the true alphas are zero and that the deviations from zero arise through chance (with p-values that are always less than 0.1% even when we allow for the clustering of returns within time periods). The evidence of nonzero alphas for the decile portfolios strengthens when we extend the analysis to the 1927–2005 period.

We find that there are three main causes of these nonzero alphas. First, the Fama-French methodology equal-weights the 2×3 size-by-BM portfolios when constructing the small-minus-big (SMB) and high-minus-low (HML) factors, even though these portfolios contain very different amounts of market cap. Relative to value-weighting — the approach taken by indices and (necessarily) by investors collectively — the FF approach overweights stocks in the small value portfolio, which have outperformed during our time period (see, for example, Loughran, 1997). Overweighting small value stocks exaggerates the return of the SMB factor. This causes the (value-weighted) Russell 2000, with its positive loading on SMB, to underperform its FF benchmark, and the S&P 500, with its negative loading on SMB, to outperform.[1]

Second, beginning with Carhart (1997), the Fama-French-Carhart (FFC) methodology generally uses the Center for Research in Security Prices (CRSP) value-weighted excess return as its market factor.[2] This index includes not only U.S. common stocks, but also non-U.S. firms, closed-end funds, real estate investment trusts (REITs), and other securities such as shares of beneficial interest (SBIs). These other assets dramatically underperformed U.S. common stocks from 1980 to 2005, yielding an annual Carhart alpha of -4.01% ($t = 2.67$). Thus, the CRSP market index underperforms U.S. common stocks by about 23 basis points per year. For indices that mainly hold U.S. common stocks, such as the S&P 500, this contributes to positive alphas.

Third, annual changes to the indexes also contribute to negative index alphas, especially for small-capitalization indices. For example, every year,

---

[1] Note that if one size factor and one value factor were enough to span average returns across the size-value grid, then any combination of the factors would still explain the cross-section of returns. Hence, the problem with nonzero alphas arises because one value factor and one size factor are not enough to span historical returns across the size-value grid; there seems to be a separate value effect for small and large stocks.

[2] Fama and French (1993) use only U.S. common stocks in the market portfolio. In subsequent papers they use the CRSP value-weighted index, which is the market return provided on Ken French's website.

at the end of June, Russell adds and deletes stocks from its indices according to a pre-announced formula. In anticipation of the one-time demand shock by index investors, stocks that are added to the Russell 2000 outperform stocks that are deleted in June, while the reverse occurs in July, lowering the returns on the index itself. We find that about one half of the negative alpha of the Russell 2000 occurs during June and July, suggesting the reconstitution effect also has an impact on index alphas.[3] But reconstitution is not the full story, because the Russell 2000 exhibits a negative, statistically and economically significant alpha in the other 10 months of the year as well, and we also find negative alphas for small-cap portfolios such as the S&P 600 and size decile portfolios that suffer from smaller or negligible reconstitution effects.

These sources of nonzero alphas represent a combination of *ex-ante* and *ex-post* biases. For instance, the outperformance of small value and underperformance of non-U.S. common stocks included in CRSP during our time period need not persist out-of-sample. In contrast, the underperformance of closed-end funds, whose returns reflect underwriting and management fees, might be expected to persist. However, even the *ex-post* biases we document are undesirable, because they affect performance evaluation in the commonly studied time period, and they indicate a general lack of robustness that could lead to biased alphas (in either direction) in future time periods.

We explore three remedies. The first modifies the Fama-French-Carhart methodology by value-weighting the SMB and HML factors and by limiting the market factor to U.S. common stocks. This brings the FFC methodology closer to the practices of the asset managers that it is used to evaluate. The second substitutes index-based factors (S&P 500 for the market; Russell 2000 minus S&P 500 for SMB; Russell 3000 Value minus Russell 3000 Growth for HML), which are value-weighted and exclude most of the underperforming securities included in the CRSP value-weighted index. The third adds three new factors: one captures the relative performance of midcaps, and the other two allow the value-growth effect to differ for large, midcap, and small cap stocks.

---

[3] We also investigated whether flows into index funds or institutional portfolios benchmarked to the various indices are related to benchmark alphas, but we did not find any robust associations.

We find that both modified FFC and index models reduce index alphas significantly, although for index models this is partly mechanical. When used to explain actively managed mutual fund returns, modified FFC and index models produce less out-of-sample tracking error volatility. Index models perform slightly better than modified FFC models in this latter analysis. Also, the seven-factor versions of the models yield lower (out-of-sample) tracking error volatility than the four-factor versions. Surprisingly, the improvements from adding three factors are fairly modest relative to the improvement from switching to our alternative four-factor models from the FFC model. Finally, we provide an example of how the benchmark model biases can affect conclusions about manager performance. We compare the average alphas of actively managed mutual funds in different size categories. Alphas from the FFC model suggest that small-cap managers underperform large-cap managers by -2.13 percentage points per year. This counter-intuitive result is fully reversed when we switch to any of our modified or index factor models.

This is not the first study to document nonzero alphas. In Table 9a and p. 41 of Fama and French (1993), the authors note a positive alpha in the large-growth corner of their 5×5 sort and a negative alpha in the small-growth corner. Table 8 in Chan *et al.* (2009) shows a negative and statistically significant alpha for the Russell 2000 Growth index. Both papers make subtle choices that — presumably unintentionally — minimize the problem. In their Table 1, Fama and French equal-weight their 25 stock portfolios in their presentation and F-statistics, which gives the positive alpha large-growth corner portfolio a 4% weight, despite its containing 30% of market capitalization. Chan *et al.* (2009) study a 13-year sample and therefore find a statistically significant alpha for only one of the eight Russell indices they study.

The lower power of regression-based factor models in a short time period is less of an issue for Chan *et al.* (2009), because their main focus is on characteristics-based models. Instead, our focus is on the more widely used factor models. In addition to more thoroughly documenting nonzero alphas, our main contribution is understanding their exact source. We trace the alphas to specific choices made in the construction of factors. Fortunately, these choices are easily altered, and with minimal effort, the profession can adopt alternative factor models that better measure portfolio performance.

Our contribution is methodological as well as conceptual and related to the benchmarking and pricing models of Fama and French (1993) and Carhart (1997). Sharpe's (1992) style analysis is one of the few studies using multiple benchmark indices for performance evaluation. However, it does not investigate model construction in any detail or evaluate alternative model specifications. In addition to indices, Elton *et al.* (1999) advocate the use of factors that are based on mutual fund returns. Daniel *et al.* (1997) present a nonlinear benchmarking methodology based on characteristics-matched portfolios which avoid many of the issues we document, albeit at the cost of requiring knowledge of portfolio holdings and a nontrivial amount of computation. In this paper, we focus on refining factor models that do not require holdings data, given that this approach remains popular among researchers and practitioners. Furthermore, we focus on unconditional factor models (unlike Ferson and Schadt, 1996) and only on the value and size dimensions. This is, again, to keep the scope of our study manageable and to make it relevant for the large number of researchers and practitioners who evaluate portfolio performance using the unconditional four-factor FFC model.

This paper proceeds as follows. Section 2 discusses the criteria for judging a benchmark model and how they differ from those that are used for pricing models. Section 3 explains the data sources, including the basics of the most common benchmark indices. Section 4 presents evidence on benchmark index alphas under the Carhart model, and investigates the sources of those alphas. Section 5 presents alternative factor models, examines the common variation in returns explained by various factor models, and explores how the choice of the model affects conclusions about the relative performance of managers in different styles. We present our conclusions in Section 6.

2 Defining a Good Benchmark Model

We define a *benchmark portfolio* as a passively managed portfolio with factor exposures similar to the portfolio whose performance we are evaluating. A *factor* is any excess return that could be used in constructing a benchmark. We define an *index* as a commonly known and used benchmark such as the S&P 500. A benchmark portfolio could simply be an index, or it could be a combination of multiple factors; for example, the Carhart four-factor

model produces a benchmark portfolio that is a weighted combination of four factor portfolios.

Criteria for defining a "good" benchmark model for portfolio performance evaluation are not identical to those of a good pricing model, even though pricing models can also be used as benchmark models. A pricing model should be the simplest possible model that explains the cross-section of expected stock returns. Asset pricing theory suggests that expected returns should be a linear function of betas of the portfolio with respect to one or more systematic risk factors. Empirically motivated factors, in principle, could be derived from any stock characteristic that is believed to predict returns out-of-sample.

A benchmark model, in contrast, should provide the most accurate estimate of a portfolio manager's added value relative to a passive strategy. This implies that a benchmark model should include the pricing model so that the manager does not get credit for simply bearing more systematic risk. A benchmark model can also include non-priced factors to reduce noise in alpha estimates, or can even encompass well-known anomalies. For example, if momentum has historically produced excess returns, but whether it will do so once it is widely known is in question, then controlling for past exposure to momentum in a benchmark might be justified. Even including industry risk in the benchmark might be warranted when evaluating the abnormal performance of a fund manager with a persistent tilt toward a particular industry, regardless of the fact that industry risk should presumably not be priced *ex-ante* and therefore should not be included in a pricing model.

In this spirit, Fama and French (1993) propose two bond market factors, despite their long-term risk premia being close to zero, both because they explain significant time-series variation in returns and because their risk premia may vary over time. Furthermore, Pastor and Stambaugh (2002) show that including non-priced factors in a benchmark model helps estimating alphas, even if we know the true *ex-ante* pricing model. However, most academics have chosen to use the Fama-French three-factor model and the Carhart four-factor model for benchmarking applications.

To determine how well a model performs as a benchmark for money managers, we test several properties. First, a new model should track the time series of portfolio returns better than old models, and produce lower out-of-sample tracking error volatility. This also means that the factors should capture common variation in returns, which is a necessary condition

for a nonzero factor premium in the Arbitrage Pricing Theory. Second, a model should explain the cross-section of average returns well, without generating significant alphas across large segments of the market such as large caps or small caps. This should hold for test assets such as size and book-to-market-sorted portfolios as well as for a cross-section of mutual funds, unless it is plausible that the average managerial skill varies from large positive to large negative values across market segments or styles.

Unfortunately, no model with a reasonable number of factors is likely to span the entire cross-section of average returns. Even across the two dimensions of size and value we would need more than three factors. For example, Fama and French (1993) report significant alphas for large growth and small growth portfolios using their three-factor model. To make comparisons across models, we must therefore decide how to weight these pricing errors across portfolios. In other words, because we cannot price everything correctly, we must identify whether we care about pricing some portfolios (such as the ones with greater market value) more than other portfolios (such as the ones with very little market value).

In this paper we work with three kinds of test assets: (1) common benchmark indices such as the S&P 500 and Russell 2000, (2) other passive portfolios such as CRSP size deciles, and (3) portfolios of actively managed U.S. equity mutual funds. The common theme in our tests is that we try to price real-world investment portfolios rather than portfolios that are narrowly defined, have little market capitalization, or have significant time variation in their composition. For example, we consider it more important to price the S&P 500 correctly than to price a portfolio of illiquid microcap value stocks with a trivial amount of market capitalization.

The question of which factor model is the best one for performance evaluation is too broad to address in this paper. Instead, we focus on a narrower question: along the two dimensions of size and value — the most common ways for both academics and practitioners to slice the equity market into different investment styles — what are the problems with existing academic performance evaluation models, and how could they be improved while maintaining the same basic structure? We build directly on the contribution of Fama and French (1993), who introduced the original size and value factors. Understanding these problems with size and value benchmarks is a prerequisite for addressing the broader question of the best factor model in all dimensions. In addition, our proposal to use benchmark indices

themselves as factors relates to industry practice. In contrast to the academic literature, practitioners generally compare money managers against their self-declared benchmark indices such as the S&P 500 or Russell 2000. The mere subtraction of the benchmark index return may oversimplify performance evaluation, so we make this approach more flexible by using a set of benchmark indices as factors for benchmarking purposes.

3 Data

3.1 Benchmark Indices

In our study, we include all non-specialized U.S. equity benchmark indices that are commonly used by practitioners. This covers a total of 23 indices from three index families: Standard and Poor's, Frank Russell, and Dow Jones Wilshire. Our data, obtained directly from these three index providers, covers monthly and daily index returns, including dividends, as well as month-end index constituents. The main S&P indices are the S&P 500, S&P MidCap 400, and S&P SmallCap 600. The S&P 500 is the most common large-cap benchmark index, consisting of approximately the largest 500 stocks. It is further divided into a growth and value style, with equal market capitalization in each. The S&P 400 and S&P 600 consist of 400 midcap and 600 small-cap stocks, and they are also further divided into separate value and growth indices. From the Russell family, we select 12 indices: the Russell 1000, Russell 2000, Russell 3000, and Russell Midcap indices, and the value and growth components of each. The Russell 3000 covers the largest 3,000 stocks in the U.S., and the Russell 1000 covers the largest 1,000 stocks. The Russell 2000 is the most common small-cap benchmark, consisting of the smallest 2,000 stocks in the Russell 3000. The Russell Midcap index contains the smallest 800 stocks in the Russell 1000. Finally, we include the Wilshire 5000 and Wilshire 4500. The Wilshire 5000 covers essentially the entire U.S. equity market with about 5,000 stocks in 2004, and peaking at over 7,500 stocks in 1998. The Wilshire 4500 is equal to the Wilshire 5000 minus the 500 stocks in the S&P 500 index, which makes it a mid- to small-cap index.

Since 1998, all mutual funds have had to report a benchmark index to the Securities and Exchange Commission (SEC). Table 1 indicates the popularity of each index, and shows the self-reported benchmark indices for U.S. all-equity mutual funds in January 2007. The most common self-declared

benchmark index is the S&P 500. The Russell 2000 is the second-most popular benchmark, and its value and growth components are also relatively popular. The most common self-declared midcap index is the S&P 400, although the Russell Midcap group of indices is collectively more popular.[4] Wilshire indices are less common in terms of the number of funds but they each have a nontrivial amount of assets benchmarked to them. S&P indices do not cover all stocks, due to S&P's relatively tight selection criteria on profitability and other firm characteristics. Additionally, the market cap boundaries of each S&P index are very flexible, as market cap is only one of S&P's selection criteria. In contrast, Russell indices cover virtually their entire target universe and use strict market cap cutoffs.[5]

3.2 Other Data Sources

All stock data are from CRSP, supplemented with accounting data from Compustat. Mutual fund data items are primarily from CRSP, with the exception of quarterly holdings data from Thomson Financial, self-reported benchmark index data from Morningstar, and daily fund returns before 2001 from a survivorship-free database that was originally obtained from the Wall Street Web and used by Goetzmann *et al.* (2001). The CRSP and Thomson Financial mutual fund databases are matched by using MFLINKS. Following the screening procedure in Cremers and Petajisto (2009), we pick a sample of U.S. all-equity mutual funds with at least $10 million in assets to address issues like incubation bias. Fama-French factor and portfolio data are from Ken French's website.

---

[4] The Russell style indices have recently become more common benchmarks than the S&P style indices, whereas the S&P 500 style indices used to be more popular in the 1990s. Boyer (2006) provides more details on the S&P 500 style indices.

[5] These points are illustrated in Figure 1 of the Online Appendix, which shows the fraction of ordinary common stock of U.S. firms covered by the most common indices as a function of market capitalization. Each month and for each market cap rank n, we compute the fraction of the neighboring 20 stocks (i.e., stocks with market cap ranks from $n - 10$ to $n + 10$) that are in the index. The figure reports the average index membership density from 1996 to 2005. We do not see discrete steps at 1,000 and 3,000 because we average across market cap rankings throughout the year, whereas Russell updates its indices only once a year.

Index	Number of mutual funds	Mutual funds assets ($M)
S&P 500	1,318	2,130,000
Russell 2000	251	214,712
Russell 1000 Growth	180	162,710
Russell 1000 Value	177	249,537
Russell 2000 Growth	132	48,579
Russell Midcap Growth	107	73,563
Russell 2000 Value	106	65,066
S&P 400	74	102,241
Russell Midcap Value	62	85,629
Russell 1000	53	56,660
Russell 3000	48	43,344
Russell Midcap	35	23,260
Russell 3000 Growth	31	67,130
S&P 600	27	14,326
Russell 3000 Value	26	63,722
Wilshire 5000	20	114,092
S&P 500 Value	8	6,307
Wilshire 4500	5	16,254
S&P 500 Growth	5	345
S&P 400 Value	4	10,869
S&P 400 Growth	3	192
S&P 600 Value	3	181
S&P 600 Growth	2	57

Table 1. The most common benchmark indices.

Description: For each index, the second column is the number of actively managed U.S. all-equity mutual funds reporting the index as their primary benchmark in January 2007. The last column is the sum of total net assets across all such funds. The data source is Morningstar. Some funds have a missing primary benchmark in the database.

Interpretation: The table points out that the S&P 500 and Russell 2000 (together with its style components) are the most popular benchmark indices.

4 Alphas of Benchmark Indices

4.1 Baseline Results

Table 2 presents estimates of Carhart and Fama-French alphas for the major Russell, S&P, and Wilshire indices from 1980 to 2005.[6] Alphas are positive and statistically significant for the general and growth versions of the large-cap indices (the Russell 1000 and S&P 500) and negative and statistically significant for the general and growth versions of the small-cap indices (the Russell 2000 and S&P 600). As expected, the alpha for the Wilshire 5000 is close to zero, because it approximates the CRSP value-weighted index (which is included as a factor in the Carhart model). Index alphas are similar for the Fama-French and Carhart models, reflecting generally minor loadings on the momentum factor, which consequently does not play any role in our analysis. Furthermore, an F-test of all index alphas jointly being equal to zero produces a p-value below 0.0001%. This means that index alphas are jointly statistically significant at any reasonable level, and highlights again that the nonzero alphas are a problem for more than a few indices.

Figure 1 plots cumulative Carhart alphas for the growth, value, and general versions of the S&P 500 and Russell 2000. The plots begin at January 1980 or the inception date (whichever is later), with the exception of the S&P 500, which we extend back to January 1961 (the beginning of the CRSP Mutual Funds dataset, and thus of many performance evaluation studies). The S&P 500 exhibits a positive alpha during the 1960s, 1980s, and 1990s, and an approximately zero alpha during the 1970s and after 2000.[7] The negative alpha of the Russell 2000 is fairly steady throughout the sample. For both indices, nonzero alphas are significantly more pronounced for the growth versions.

---

[6] We use a sample period back to January 1980 when possible. For some indices (see the footnote to Table 2 for a list), the first available return data are from a later date. The Russell indices were introduced in January 1984, and returns from 1980 to 1983 were calculated by Russell based on a back-casting of their index construction rule (which is mechanically based on market capitalization). Following most of the recent literature, we calculate our alphas in-sample, estimating factor weights over our entire sample period. In unreported results, index alphas estimated using betas from a trailing 60-month window are qualitatively similar.

[7] Since writing the first draft of this paper in 2007, we have extended the results in Figure 1 and Table 2 to the end of 2010. Including this period reinforces our conclusions from 1980–2005. In particular, large-cap indices continued to have more positive FFC alphas than small-cap indices, and this continued to be particularly true among the growth sub-indices. Versions of Figure 1 and Table 2 that extend to 2010 are available in the Online Appendix.

	Carhart Alpha			Fama-French Alpha		
Main Index	Growth	All	Value	Growth	All	Value
Russell 3000	1.05	0.18	−0.58	1.13	0.04	−1.30
	(1.96)	(0.95)	(1.03)	(2.43)	(0.24)	(2.28)
Russell 1000	1.53	0.47	−0.48	1.60	0.33	−1.19
	(2.60)	(2.60)	(0.86)	(3.11)	(1.86)	(2.10)
Russell Midcap	1.61	0.17	−0.52	1.98	0.04	−1.29
	(1.39)	(0.19)	(0.62)	(1.65)	(0.06)	(1.36)
Russell 2000	−3.41	−2.41	−1.25	−3.27	−2.53	−2.09
	(3.95)	(3.21)	(1.23)	(3.81)	(3.60)	(2.11)
S&P 500	1.82	0.82	−0.35	2.40	0.72	−1.61
	(2.91)	(2.78)	(0.72)	(3.65)	(2.60)	(2.57)
S&P Midcap 400	0.40	1.36	0.89	−0.94	1.41	2.41
	(0.20)	(1.31)	(0.46)	(0.44)	(1.34)	(1.13)
S&P Smallcap 600	−3.09	−2.64	−1.61	−1.53	−2.62	−2.91
	(1.32)	(2.26)	(1.02)	(0.72)	(2.41)	(1.78)
Wilshire 5000		0.05			0.05	
		(0.43)			(0.41)	
Wilshire 4500		−0.96			−0.49	
		(1.40)			(0.75)	
P-value of joint significance test for 26 indices		<0.000001			<0.000001	

Table 2. Alphas of benchmark indices.

Description: This table shows the Carhart four-factor alphas as well as the Fama-French three-factor alphas for common benchmark indices. Alphas are computed from monthly data. The numbers shown are expressed in percent per year, with absolute values of heteroskedasticity-robust t-statistics in parentheses. The sample period is January 1980 to December 2005, except for the following indices whose return data begin later: S&P 400 (2/1981), Wilshire 4500 (1/1984), S&P 600 (3/1984), and the Growth and Value components of the Russell Midcap (2/1986), S&P 400 (6/1991), and S&P 600 (1/1994). The variance-covariance matrix used in the joint significance test allows for clustering of index returns within time periods.

Interpretation: The table points out that many common indices have significant nonzero Carhart and Fama-French alphas, which is surprising for diversified and entirely passive portfolios. The test of joint significance strongly rejects the null hypothesis of zero alphas.

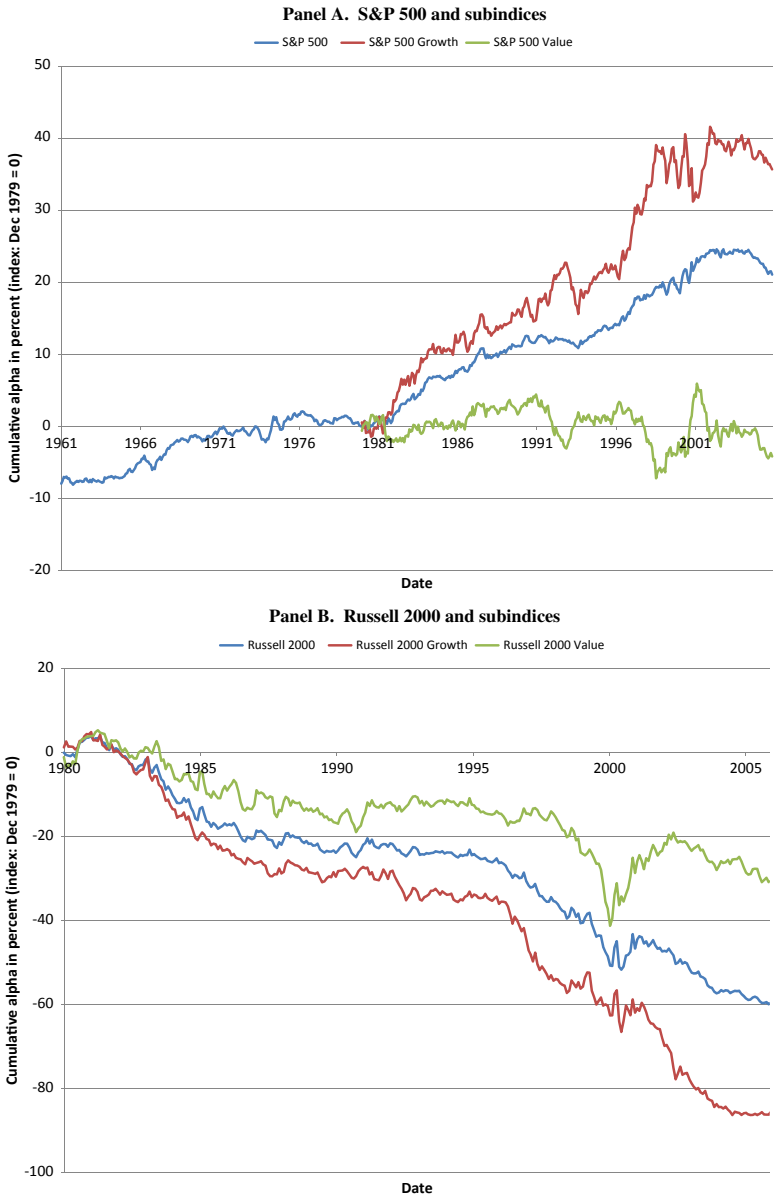

Figure 1. Abnormal returns of S&P 500 and Russell 2000.

Description: This figure plots the cumulative abnormal alpha from the Carhart model, estimated insample for the time period 1980-2005 (except for the S&P 500, which is 1961-2005).

Cumulative alphas are expressed in log percentage points and normalized to zero at the beginning of 1980.

Our main point does not depend on whether these patterns exist in every time period. Having both economically and statistically significant alphas in any time period potentially indicates a lack of robustness for the benchmark model, and it is particularly troubling in the 1980–2005 time period because this period is so commonly used by researchers.

4.2 Sources of Index Alphas: Factor Construction

We now reexamine the methodological choices of the standard Fama-French model and evaluate how they contribute to the index alphas. On page 9 of Fama and French (1993), they note that the choices made in constructing their factors "are arbitrary . . . and we have not searched over alternatives." This reduced the temptation of data mining. This is an important concern for us as well; in proposing or recommending alternative choices, we are guided by our effort to mimic the choices made by funds and practitioners. Specifically, we examine four methodological choices in the construction of the benchmark model: (1) the universe of assets included in the market factor, (2) the weighting of component portfolios when constructing factors, (3) the imposition of a common value factor for small and large stocks, and (4) the boundaries between size and BM categories. In each case, we propose alternative choices that are more consistent with the construction of the benchmark indices and real-world portfolios. We find that these alternative choices lead to factor models that approximate the mix of stocks held by comparable indices more closely, and, individually and collectively, reduce index alphas and their variance.

4.2.1 Definition of the Market Portfolio

For their market proxy, Fama and French (1993) use a value-weighted portfolio of the stocks in their size and BM portfolios, plus stocks with negative book equity. They include common stocks of U.S.-headquartered and listed firms — CRSP share codes 10 and 11 — that have a sufficiently long history,[8] and therefore exclude new issues. Carhart (1997), and most of the subsequent literature, use the CRSP value-weighted index instead. This

[8] This means that Compustat and CRSP data for the firm must have started 3.5–4.5 years and 0.5–1.5 years earlier, respectively, depending on the month.

includes all U.S.-headquartered and listed common stocks, as well as closed-end funds, REITs, foreign firms with primary listings in the U.S., and other asset types such as certificates, shares of beneficial interest, and units.[9] This is also the market return that researchers commonly obtain from Ken French's website.

We find that the choice of included securities in the market proxy significantly affects measurements of risk-adjusted returns. Table 3 reports Carhart alphas for the different components of the CRSP value-weighted index. This index has an alpha of exactly zero by construction because it is included as a factor in the model. U.S. common stocks — CRSP share codes 10 and 11 — collectively have an alpha of 23 basis points (bp) per year over the 1980–2005 period. This is explained by the underperformance of other securities such as foreign firms and closed-end funds, which have a surprisingly low Carhart alpha of -4.01% ($t = 2.67$) per year. The stocks included in the Fama-French (1993) market factor, which are those included in their size-BM-sorted portfolios, have an alpha of 51 bp per year. This indicates underperformance of stocks with insufficient data or negative book value, which is also consistent with the general long-term underperformance of IPOs.[10]

One might view non-zero alphas of these additional categories of assets as anticipatable *ex ante* (for example, for closed-end funds of which the returns reflect underwriting and management fees), or as merely an *ex post* outcome. Either way, it is appropriate to benchmark managers against the returns of securities in their choice set. The Carhart model is most often used as a benchmark for domestic non-specialized equity mutual fund portfolios. We therefore use the holdings of these portfolios and their self-declared benchmark indices as a guideline for what to include in the market factor (Table 3). New issues are included in these portfolios, while closed-end funds, foreign firms, and assets such as shares of beneficial interest are excluded from the indices, and are held at much lower rates by funds, if at all. Foreign firms are less likely to be included in domestic equity indices or funds. REITs are the closest call as they are held by the benchmark indices

[9] American Depository Receipts (ADRs) are the only securities included in the CRSP stock file but excluded from the CRSP value-weighted index.

[10] See Ritter (1991) for the long-term IPO performance, and also Barber and Lyon (1997, p. 342), who discuss the associated reverse problem of the "new listing bias, which arises because... sampled firms generally have a long post-event history of returns, while firms that constitute the index typically include new firms that begin trading subsequent to the event month."

Group	Share codes (from largest to smallest market cap)	Average share of CRSP VW	Carhart alphas		Percent of capitalization held by:			
			Percent per year	t-stat	S&P 500	Russell 3000	Wishire 5000	Equity funds
U.S. common stocks	11, 10	92.68%	0.23	(2.00)	77.4	97.0	98.9	10.12
Subset included in FF portfolios	11, 10	87.87%	0.51	(2.68)				
Subset not included in FF portfolios	11, 10	4.81%	−2.74	(1.66)				
All other securities in CRSP VW index	See below	7.32%	−4.01	(2.67)	12.4	14.8	24.0	4.88
Non-US stocks, units, and SBIs	12, 72, 42	4.76%	−3.74	(2.00)	14.6	0.9	12.3	5.57
Closed-end funds	14, 44, 15, 74, 24	1.06%	−1.65	(1.02)	0.0	0.1	0.1	0.09
REITs	18, 48	0.74%	−0.75	(0.37)	21.3	97.2	99.8	8.35
Other (certificates, SBIs, units)	71, 23, 73, 70, 41, 21, 40, 20	0.76%	−3.39	(1.85)	0.0	0.5	12.4	0.79
CRSP VW index	All except ADRs	100%	0.00	(0.00)	69.6	87.0	89.8	9.49
ADRs (excluded from CRSP VW)	31, 30	3.31%	4.25	(1.55)	0.0	0.0	0.0	

Table 3. Four-factor alphas by CRSP share code, 1980–2005.

Description: This table aggregates the share codes reported in CRSP into groups. The CRSP value-weighted (VW) index consists of all share codes except ADRs. The table reports the average share of the CRSP VW index accounted for by each group from 1980-2005, along with their four-factor Carhart alphas. The Carhart alpha of the CRSP VW index is, of course, zero by construction. The table also reports, based on December 2004 data, the case of each groups capitalization that is a member of three indices (the S&P 500, Russell 3000, and Wilshire 5000) and the share that is reported as holdings by U.S. equity mutual funds on SEC form 12D. Among U.S. common stocks, Fama and French (1993) portfolios exclude new stocks and negative BM stocks. The CRSP VW index is used as the market portfolio on Ken Frenchs website. Absolute values of t-statistics from robust standard errors are in parentheses.

Interpretation: The table highlights that the CRSP VW index includes a nontrivial investment in securities other than U.S. stocks and that those other securities have a statistically significant and economically large negative Carhart alpha, thus inducing an offsetting positive alpha for U.S. stocks.

and by equity mutual funds. The funds do however represent a slightly smaller fraction of shareholders in REITs than in U.S. firms. For this reason, we exclude them from the market factor, but as their inclusion affects the average return of the market proxy by less than 1 basis point per year, results would therefore be very similar if they are included.

Overall, these results indicate that the CRSP value-weighted market portfolio is a downward-biased benchmark for portfolios consisting of U.S. common stocks, even within our relatively long time period of two and a half decades. Instead, it would be more appropriate to benchmark actively managed portfolios that only hold U.S. equities with a market portfolio consisting of U.S. equities.[11]

4.2.2 Equal-Weighting in Fama-French Factors

The second Fama-French methodological choice involves the weighting of stocks in constructing factors. Fama and French (1993) construct factors that capture the relative performance of small and value stocks by using the following procedure: they sort U.S. common stocks into six value-weighted portfolios based on whether a stock's market capitalization is "Big" (above the NYSE median) or "Small" (below the median), and whether its BM ratio is "High" (top three deciles), "Medium" (middle four deciles), or "Low" (bottom three deciles). They then equal-weight across these six portfolios: their small-minus-big (SMB) factor is (Small-Low + Small-Medium + Small-High)/3 − (Big-Low + Big-Medium + Big-High)/3, and their high-minus-low-BM (HML) factor is (Small-High + Big-High)/2 − (Small-Low + Big-Low)/2 as illustrated in Panel B of Table 4. Fama and French exclude from the six portfolios stocks with negative book equity or with no book equity data available for the fiscal year ending in the prior calendar year, and so these stocks receive zero weight in their factors.

Panel A of Table 4 reports the average share of the CRSP market index as represented by Fama and French's 2×3 portfolios as well as their average excess returns. Panel A also shows two portfolios of stocks that are excluded

[11] Figure 3A in the Online Appendix plots the cumulative difference between the value-weighted return of all U.S. common stocks in CRSP and the CRSP-VW return (which, as discussed, includes other types of securities). The apparent downward bias in the CRSP-VW return is most pronounced in the 1980s and 1990s, which is also when the S&P 500 exhibits a positive alpha.

Panel A: Market portfolio weights and component returns (%)

	MktRf weights						Average excess return per year				
	None	Gro	Med	Val	All		None	Gro	Med	Val	All
Big	7.8	42.6	25.5	11.1	86.9	Big	5.92	7.61	8.62	9.20	7.72
Small	4.2	3.5	3.4	2.0	13.1	Small	6.47	4.85	11.77	13.21	8.29
All	12.0	46.1	28.9	13.0	100.0	All	5.87	7.20	8.95	10.02	7.64

Panel B: Fama-French factor portfolio weights (%)

	SMB						HML				
	None	Gro	Med	Val	All		None	Gro	Med	Val	All
Big	0.0	−33.3	−33.3	−33.3	−100.0	Big	0.0	−50.0	0.0	50.0	0.0
Small	0.0	33.3	33.3	33.3	100.0	Small	0.0	−50.0	0.0	50.0	0.0
All	0.0	0.0	0.0	0.0	0.0	All	0.0	−100.0	0.0	100.0	0.0

Panel C: Target portfolio weights vs. their three-factor benchmark weights (%)

	Target portfolio: Size decile 10						Benchmark portfolio: $0.967 \times$ MktRf $- 0.318 \times$ SMB $- 0.086 \times$ HML				
	None	Gro	Med	Val	All		None	Gro	Med	Val	All
Big	0.0	60.0	29.2	10.8	100.0	Big	7.5	56.1	35.2	17.0	115.8
Small	0.0	0.0	0.0	0.0	0.0	Small	4.1	−2.9	−7.3	−13.0	−19.1
All	0.0	60.0	29.2	10.8	100.0	All	11.6	53.2	27.9	4.0	96.7
	Target portfolio: Size decile 4						Benchmark portfolio: $1.055 \times$ MktRf $+ 0.799 \times$ SMB $+ 0.226 \times$ HML				
	None	Gro	Med	Val	All		None	Gro	Med	Val	All
Big	0.0	0.0	0.0	0.0	0.0	Big	8.2	7.0	0.3	−3.7	11.8
Small	0.0	40.7	40.5	18.7	100.0	Small	4.5	19.1	30.2	40.0	93.8
All	0.0	40.7	40.5	18.7	100.0	All	12.7	26.1	30.5	36.3	105.5

Table 4. Comparing actual portfolios with their Fama-French benchmarks.

Description: This table shows the benchmark portfolio holdings implied by the three-factor Fama-French model. These holdings are contrasted with the true holdings of the target portfolios we are trying to explain. As target portfolios, we pick the FF size deciles 10 (large-cap stocks) and 4 (small-cap stocks) within the 100 FF portfolios, because they represent the typical S&P 500 and Russell 2000 constituent stocks, respectively. Panels A and B show the portfolio weights of the three FF factors, together with the excess return on the 2×3 portfolio components. The market excess return over the one-month T-bill return is denoted MktRf. Because the market factor includes CRSP securities that are not part of the 2×3 FF grid, we include these stocks in a separate "None" column. Panel C shows the true weights that each of the two target portfolios (size deciles) have on the extended 2×4 grid, alongside the weights implied by the three-factor model. The implied weights can be derived from the three-factor betas multiplied by the factor portfolio weights; the regression betas are shown above the implied portfolio weights. The time period is from 1980 to 2005.

Interpretation: The table points out that the benchmark portfolio implied by the FF three-factor model can deviate significantly from the portfolio it is used to benchmark, as measured by exposures to the 2×4 size-value portfolios. The difference in exposure to the small value portfolio matters particularly because of its high average return.

by the Fama-French factors but still included in the CRSP index. Just like in Fama and French (1993), two things are apparent: first, the growth portfolios have much more market capitalization than the value portfolios, and second, the best performance by far has been exhibited by the Small Value (or Small size — High book-to-market) portfolio.

We now show how this weighting scheme induces non-zero alphas in size-sorted portfolios, by comparing value-weighted small-cap and large-cap portfolios to benchmarks with too much or too little exposure to Small Value. While we will turn to indices shortly, we begin by examining two size decile portfolios: the Fama-French size decile 10 portfolio, which contains the large stocks typical of the S&P 500 index, and the size decile 4 portfolio, which contains the small stocks typical of the Russell 2000 index. The fitted regression of either portfolio on the Fama-French factors produces the three-factor benchmark portfolio, where the alpha is the difference in return between the target portfolio and its benchmark. If the benchmark portfolio has the same broad category exposures as the target portfolio, the alphas are likely to be zero; if the two differ significantly, this may be, but does not have to be, a source of nonzero alpha. We conduct the analysis for the Fama-French three-factor model to keep it more transparent, but the mechanism is virtually identical for the Carhart model with the added momentum factor.

The left-hand side of Panel C shows the weights that size decile 10 — large stocks — has on the 2×4 grid. The right-hand side of the panel shows the regression coefficients when the return on this portfolio is regressed on the returns on the Fama-French factors: the negative beta on SMB is expected, but the nonzero beta on HML is more surprising. Below the factor betas, we see the 2×4 portfolio weights in the benchmark portfolio that is implied by the three-factor model (0.967*MktRf − 0.318*SMB − 0.086*HML). The 2×4 weights of the size decile 10 target portfolio differ from the benchmark weights, particularly in small caps. The target portfolio has a zero weight on small caps, while the benchmark portfolio has a large and negative weight of −19.1%. Two-thirds of this difference comes from heavy underweighting of small value stocks, which have performed very well (Panel A). This significant underweighting of small value stocks contributes to poor performance by the benchmark and thus a positive alpha relative to this benchmark for this target portfolio.

Why does the decile 10 benchmark portfolio get such a large underweight on small value stocks? Because the market beta is approximately 1, we start the benchmark portfolio with the market weights in Panel A. As previously discussed, SMB places equal weights on all six component portfolios (Panel B), so it will reduce the weight on small value stocks, that have a market weight of 2.0%, too much compared to small growth stocks, that have a market weight of 3.5%. Furthermore, a large negative beta on SMB will have too much weight on large value stocks while not increasing the weight on large growth stocks enough. To reduce this overweight on large value, the model produces a negative beta on HML. This comes however at the cost of reducing the weight on small value even more, producing a 13% underweight.

The small stocks in size decile 4 exhibit the opposite effect. When the returns of this portfolio are regressed on the three-factor model, the market beta is again approximately 1, but SMB and HML betas are positive. The equal-weighting of SMB implies that the large positive SMB beta produces an overweight in small value and an underweight in small growth stocks. Furthermore, the SMB weights generate a considerable growth bias in large stocks: about +18% weight in large growth and −15% weight in large value. A positive HML loading is needed to offset this growth tilt, but it comes with the cost of increasing the small-cap value bias even more. As a result, the benchmark portfolio has a 40% weight on small value compared to the target portfolio's 19% weight, and it has the opposite weights on small growth. Given the performance record of small value relative to small growth stocks (Panel A), this value tilt in the three-factor benchmark significantly contributes to a negative alpha on the target portfolio.

A simple way to address this problem is to value-weight the SMB component portfolios within the size groups. We conclude that an equal-weighted SMB distorts alphas in two ways. First, overweighting value creates a high average return for an equal-weighted SMB factor. Second, the equal-weighted SMB factor distorts portfolio weights in large stocks in a way that induces an offsetting HML loading. Taken together, these effects lead to an underweighting of small value stocks in the benchmark for large-cap portfolios and an overweighting of small value stocks in the benchmark for small-cap portfolios. This, in turn, contributes to a positive alpha for large stocks and a negative alpha for small stocks. Having a value-weighted SMB factor avoids both problems. We quantify this effect in Section 4.3.

4.2.3 Single Value Factor across Size Groups

One might still wonder why the weighting of the factor portfolios matters, and, in particular, why betas do not fully adjust according to how the factors are constructed. For example, if we tilt the HML factor toward a very high exposure to small value stocks, why does this not simply produce lower HML betas and thus leave alpha estimates unaffected? The reason is that three factors are not enough to span the average returns across the size-value grid; the value premium has been much stronger in small caps than large caps across our sample period, so a single value factor cannot capture it. Because of this incomplete spanning, one factor model is not just a rotation of another factor model; instead, it does matter how the factors are selected. For example, as the returns illustrate in Panel A of Table 4, the outperformance of value stocks over growth stocks is much more pronounced among small stocks ($13.21 - 4.85 = 8.36\%$ per year) than large stocks ($9.20 - 7.61 = 1.59\%$ per year). Nevertheless, Fama and French and the subsequent literature apply a single value factor (HML) that equally weights the outperformance of value stocks among small and large stocks. Using a model that forces the large-cap and small-cap value effects to be equal is likely to generate positive alphas for small value and large growth portfolios and negative alphas for small growth and large value portfolios, and this is indeed what we find in Table 2.

While the Arbitrage Pricing Theory (APT) (Ross, 1976) predicts that returns should be linearly related to factors, the APT does not rule out separate value factors for large and small stocks. Indeed, the industry practice of focusing portfolios on a particular capitalization range suggests that a decoupling of the large and small-cap value effects could make sense. As a result, we experiment with models that allow for separate big- and small-stock HML factors (BHML and SHML, respectively; see also Moor and Sercu, 2006).[12]

4.2.4 Boundaries between Size and Value Groups

The fourth methodological choice we revisit is Fama and French's decision to partition stocks into two size categories (big and small) and three or

[12] Figure 3B in the Online Appendix plots the cumulative difference in the log returns of the SHML and BHML factors since 1961. The time periods in which Small Value differentially outperforms match the time periods in which the S&P 500 exhibits a positive alpha (the 1960s, and especially the 1980s and 1990s).

four value categories (low, medium, or high BM, and none). In contrast, the industry practice has been to partition stocks into three or four size categories (large, mid, small, and micro) but only two value categories (growth and value, with some indices and portfolios including both). This practice is reflected in both the Russell and S&P families of indices.

The S&P 500 primarily includes stocks from NYSE size deciles 9 and 10, and midcap stocks are drawn mostly from deciles 6–8. The Russell 2000 includes stocks from size deciles 2–5, and the microcaps, included only in the Wilshire 5000, are primarily in decile 1. The growth components of the indices include stocks from only the two to three lowest BM deciles, while stocks in the other seven to eight BM deciles are usually in the value index. This is because the indices construct the growth and value components so that they evenly divide the market cap of the index, whereas the Fama-French decile cutoffs weight stocks equally and are based only on NYSE stocks, which tend to have a value bias relative to Nasdaq stocks.

In Figure 2 of the Online Appendix, we report the SMB and HML betas of a 10×12 matrix of size-value portfolios — the standard Fama-French 10×10 size-BM-sorted matrix with additional columns for U.S. common stocks omitted from the standard matrix (labeled "N" for no book-to-market data, these include new listings) and other securities included in the CRSP-VW index (labeled "O" for other). The betas yield three interesting patterns.[13] First, only the largest cap decile is clearly negatively correlated with SMB whereas the midcaps (size deciles 6–8) are positively correlated with SMB despite being included in Big stocks, which should mechanically induce a negative correlation. Second, BM deciles 4–9 (Medium and High in the Fama-French scheme) are all positively correlated with HML. Third, the N (no BM) column has a modest negative correlation with HML. One could argue, based on these correlations, that midcaps should be included with Small rather than Big stocks, Medium-BM stocks should be included with High-BM stocks, and the None portfolio of stocks should be included with Low-BM stocks.

Given these results, we suggest modifications to make the academic partitions more closely reflect the industry approach. The first is to divide Big stocks (NYSE size deciles 6–10) into large-cap (deciles 9 and 10) and midcap stocks (deciles 6–8). The second modification is to include Medium-BM stocks with High-BM stocks, which also results in the share of capitalization treated as Value and Growth being closer to the 50-50 split used in the

[13] Results are not reported in order to save space but are available on the journal's website.

industry. We do not include stocks in the None portfolios with the Low-BM stocks because some of these stocks can be characterized as extreme value stocks (for example, those in financial distress with negative book equity), although including them makes little difference to the results that follow.

4.3 Impact of Alternative Models on Benchmark Alphas

4.3.1 Alphas of Common Benchmark Indices

In this section, we turn from analyzing size decile portfolios to popular benchmark indices and examine how alternative choices in constructing factors affect the index alphas as well as their implied loadings on size-BM portfolios. Panel A of Table 5 contains the results for the S&P 500 and Panel B for the Russell 2000.[14] Each panel estimates several alternative models and calculates the weights implied by the resulting betas on a 3×4 set of size-BM portfolios (Large, Mid, and Small size; Low, Medium, High, and No BM).[15] These implied weights are then compared with the weights estimated using a "flexible model" (which includes each of the 12 portfolios as a factor). It is also compared with a "flexible Non-Negative Least Squares (NNLS)" version that restricts all weights to be nonnegative, and with the actual percentage of the index accounted for by each portfolio as calculated from holdings data. This comparison helps identify instances in which the structure of the factor model leads to a mismatch between the model-implied weights on the 3×4 portfolios and the index's actual weights. Although such mismatches need not necessarily contribute to index alphas, for the indices examined, we find that models producing close portfolio weight matches also produce smaller index alphas.

The first column in Panel A of Table 5 estimates the standard Carhart four-factor model for the S&P 500. The second column replaces the CRSP value-weighted index (CRSP-VW) with a value-weighted average of U.S. common stocks only (share codes 10 and 11). The third column replaces the

[14] The full table (available upon request) contains nine panels, one each for the combined, growth, and value versions of the S&P 500, Russell 2000, and Russell Midcap.

[15] Each model implies a benchmark portfolio, given by the sum of the product of the Fama-French-Carhart factor portfolios and the estimated betas. This particular benchmark portfolio (the "fitted" or explained return) in turn implies specific weights on the portfolios in the 3×4 size-BM space, which can be quite different from the actual average weights of the benchmark on these portfolios (based on the flexible model including all 12 factors or the holdings).

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Share codes in market factor	CRSPVW	10/11	10/11	10/11	10/11	10/11				
SMB weighting	EW	EW	VW	VW	VW	VW				
SMB stocks included	As in FF	As in FF	As in FF	All	All	All				
Cutoff for Big stocks	50th pct	50th pct	50th pct	50th pct	50th pct	80th pct				
Size deciles included in BHML	N/A	N/A	N/A	N/A	Top 5	Top 2				
Size deciles included in SHML	N/A	N/A	N/A	N/A	Btm 5	Btm 5				
BM deciles included in H	Top 3	Top 3	Top 3	Top 3	Top 3	Top 7				
Obs	312	312	312	312	312	312	312	312	312	312
Adjusted R^2	0.9924	0.9934	0.9934	0.9939	0.9941	0.9957	0.9882	0.9882	N.M	N.M
Constant (% per year)	0.82	0.59	0.33	0.32	0.11	0.21	0.07	0.16	0.35	0.04
	(2.78)	(2.12)	(1.23)	(1.24)	(0.43)	(0.91)	(0.20)	(0.47)	(1.79)	(0.10)
UMD	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.01	−0.01
	(3.28)	(3.43)	(3.57)	(3.49)	(3.77)	(3.67)	(2.83)	(3.03)	(4.00)	(1.26)
MktRF	1.00	0.99	1.00	1.00	1.01	1.01				
	(0.14)	(0.77)	(0.01)	(0.27)	(0.91)	(2.16)				
SMB	−0.21	−0.19	−0.18	−0.19	−0.18					
	(23.88)	(23.05)	(23.80)	(24.95)	(21.38)					
Mid minus Big (MMB)						−0.21				
						(22.45)				
Small minus Mid (SMM)						−0.09				
						(6.68)				
HML	0.01	0.02	0.06	0.05						
	(0.69)	(1.87)	(5.07)	(4.54)						
BHML					0.00	−0.01				
					(0.20)	(0.72)				
SHML					0.05	0.04				
					(4.85)	(2.33)				
MidHML						0.04				
						(2.01)				

(*Continued*)

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Average weights on 3 × 4 portfolios implied by models							Flexible	NNLS	Actual	Market
Large_Low-RF	0.451	0.440	0.453	0.457	0.477	0.524	0.541	0.544	0.507	0.396
Large_Med-RF	0.285	0.279	0.278	0.278	0.276	0.295	0.247	0.245	0.278	0.230
Large_High-RF	0.153	0.152	0.139	0.136	0.116	0.125	0.122	0.130	0.112	0.098
Large_None-RF	0.012	0.012	0.012	0.014	0.014	0.015	0.016	0.016	0.009	0.012
Mid_Low-RF	0.083	0.081	0.083	0.084	0.087	−0.014	0.020	0.00	0.031	0.073
Mid_Med-RF	0.081	0.079	0.079	0.079	0.078	0.045	0.015	0.00	0.035	0.065
Mid_High-RF	0.055	0.054	0.050	0.049	0.041	0.024	0.008	0.014	0.020	0.035
Mid_None-RF	0.012	0.012	0.012	0.014	0.014	0.004	0.032	0.030	0.002	0.012
Small_Low-RF	−0.033	−0.033	−0.062	−0.045	−0.069	−0.025	−0.017	0.00	0.001	0.042
Small_Med-RF	−0.033	−0.027	−0.030	−0.019	−0.016	0.037	−0.086	0.00	0.002	0.038
Small_High-RF	−0.044	−0.032	0.011	0.013	0.042	0.022	0.066	0.00	0.002	0.023
Small_None-RF	0.023	0.023	0.023	−0.011	−0.009	0.008	0.027	0.014	0.00	0.023

Table 5. Panel A. S&P 500 alphas and betas for different versions of factor models.

Description: The Carhart model is estimated for various versions of the SMB and HML factors, and the average implied weights the model places on each of the 3 × 4 Size-Book-to-Market (BM) portfolios are calculated. This is compared with a flexible model in which the excess returns of the index are regressed on those of the 3 × 4 portfolios. Model 1 is the standard Carhart model. Model 2 excludes share codes other than 10 and 11 (U.S. common stocks) from the CRSP VW index. Model 3 replaces the equal-weighted SMB factor with one in which the Small and Big portfolios value-weight their Low-, Medium-, and High-BM components. Model 4 includes the "No or Negative BM components in Small and Big. Model 5 calculates separate HML factors for Big and Small (e.g., BHML = Big High − Big Low). Model 6 splits SMB into "Mid minus Big" (deciles 6-8 minus deciles 9-10) and "Small minus Mid." Also shown is a "Flexible" 13-factor model with all 3 × 4 portfolios and momentum, and "Flexible NNLS" where the betas on these 13 factors are constrained to be nonnegative. Absolute values of t-statistics based on Whites robust standard errors are in parentheses. The time period is from 1980 to 2005.

Interpretation: The table shows how a large fraction of the S&P 500 alpha for the Carhart model (1) is eliminated simply by using only U.S. stocks for the market (2) and by value-weighting the SMB components (3). Allowing separate value factors for small caps and large caps (models (5) and (6)) further reduces the alpha.

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Share codes in market factor	CRSPVW	10/11	10/11	10/11	10/11	10/11				
SMB weighting	EW	EW	VW	VW	VW	VW				
SMB stocks included	As in FF	As in FF	As in FF	All	All	All				
Cutoff for Big stocks	50th pct	50th pct	50th pct	50th pct	50th pct	80th pct				
Size deciles included in BHML	N/A	N/A	N/A	N/A	Top 5	Top 2				
Size deciles included in SHML	N/A	N/A	N/A	N/A	Btm 5	Btm 5				
BM deciles included in H	Top 3	Top 3	Top 3	Top 3	Top 3	Top 7				
Obs	312	312	312	312	312	312	312	312	312	312
Adjusted R^2	0.9686	0.9695	0.9838	0.9796	0.9795	0.9819	0.9862	0.9859	N.M	N.M
Constant (% per year)	−2.41	−2.66	−1.62	−1.53	−1.50	−1.61	−2.13	−2.17	−1.07	−1.23
	(3.21)	(3.64)	(2.92)	(2.44)	(2.36)	(2.83)	(4.12)	(4.16)	(2.50)	(2.40)
UMD	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	0.02	0.02	.00	−0.01
	(0.28)	(0.33)	(0.46)	(0.50)	(0.49)	(0.43)	(2.17)	(1.84)	(0.29)	(0.88)
MktRF	1.06	1.06	1.03	1.02	1.02	1.02				
	(4.34)	(4.18)	(2.97)	(2.02)	(1.88)	(1.31)				
SMB	0.80	0.82	0.81	0.81	0.81					
	(30.89)	(32.13)	(46.67)	(44.26)	(35.78)					
Mid minus Big (MMB)						0.78				
						(26.10)				
Small minus Mid (SMM)						0.70				
						(19.75)				
HML	0.20	0.21	0.06	0.09						
	(6.03)	(6.53)	(2.59)	(3.78)						
BHML					0.05	0.03				
					(1.84)	(1.02)				
SHML					0.04	0.06				
					(2.00)	(1.28)				
MidHML						0.02				
						(0.49)				

(*Continued*)

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Average weights on 3 × 4 portfolios implied by models							Flexible	NNLS	Actual	Market
Large_Low-RF	0.110	0.097	0.027	0.018	0.015	−0.043	−0.030	0.000	0.000	0.396
Large_Med-RF	0.036	0.030	0.030	0.033	0.033	0.011	−0.045	0.000	0.000	0.230
Large_High-RF	−0.020	−0.021	0.034	0.048	0.051	0.005	0.039	0.008	0.000	0.098
Large_None-RF	0.012	0.012	0.012	0.002	0.002	0.000	−0.004	0.000	0.000	0.012
Mid_Low-RF	0.020	0.018	0.005	0.003	0.003	0.082	0.120	0.062	0.040	0.073
Mid_Med-RF	0.010	0.008	0.009	0.009	0.009	0.104	0.143	0.105	0.032	0.065
Mid_High-RF	−0.007	−0.007	0.012	0.017	0.018	0.056	−0.008	0.000	0.010	0.035
Mid_None-RF	0.013	0.013	0.012	0.002	0.002	0.017	−0.027	0.000	0.007	0.012
Small_Low-RF	0.213	0.213	0.343	0.268	0.271	0.221	0.302	0.322	0.323	0.042
Small_Med-RF	0.308	0.314	0.338	0.285	0.284	0.288	0.413	0.418	0.321	0.038
Small_High-RF	0.391	0.403	0.233	0.218	0.213	0.173	0.092	0.116	0.173	0.023
Small_None-RF	0.024	0.024	0.024	0.171	0.171	0.151	0.030	0.002	0.093	0.023

Table 5. Panel B. Russell 2000 alphas and betas for different versions of factor models.

Description: The Carhart model is estimated for various versions of the SMB and HML factors, and the average implied weights the model places on each of the 3 × 4 Size-Book-to-Market (BM) portfolios are calculated. This is compared with a flexible model in which the excess returns of the index are regressed on those of the 3 × 4 portfolios. Model 1 is the standard Carhart model. Model 2 excludes share codes other than 10 and 11 (U.S. common stocks) from the CRSP VW index. Model 3 replaces the equal-weighted SMB factor with one where the Small and Big portfolios value-weight their Low, Medium, and High BM components. Model 4 includes the "No or Negative" BM components in Small and Big. Model 5 calculates separate HML factors for Big and Small (e.g., BHML = Big High − Big Low). Model 6 splits SMB into "Mid minus Big" (deciles 6-8 minus deciles 9-10) and "Small minus Mid. Also shown is a "Flexible" 13-factor model with all 3 × 4 portfolios and momentum, and "Flexible NNLS" where the betas on these 13 factors are constrained to be nonnegative. Absolute values of *t*-statistics based on Whites robust standard errors are in parentheses. The time period is from 1980 to 2005.

Interpretation: The table shows how a large fraction of the Russell 2000 alpha for the Carhart model (1) is eliminated simply by value-weighting the SMB components (3). However, some negative alpha remains.

equal-weighted SMB of the Fama-French model with a version that value-weights the High-, Medium-, and Low-BM portfolios; the fourth column also includes the No BM stocks in SMB. The fifth column replaces HML with BHML and SHML. The sixth column moves the size and BM boundaries to correspond more closely with industry practice, including Medium-BM stocks with High-BM stocks in constructing the HML factors and splitting midcaps apart from Big stocks, which involves replacing SMB with SMM (Small minus Mid) and MMB (Mid minus Big) and adding a Midcap HML factor (Mid High minus Mid Low).[16]

The alpha of the S&P 500, which is 82 bp per year in the Carhart model, declines as the models become more flexible. Replacing the CRSP-VW index with U.S. common stocks (column 2) reduces the alpha by 23 bp, or roughly the difference in the average returns of these two indices. Value-weighting SMB (column 3) decreases the alpha by another 26 bp to 33 bp per year, which is no longer statistically significant. Replacing HML with BHML and SHML (column 5) further decreases the alpha to 11 bp per year, whereas moving the size and BM boundaries (column 6) marginally increases the alpha to 21 bp. Overall, the first two steps (up to column 3) are the most important in terms of changing the alpha, and they also bring the model-implied 3×4 portfolio weights closer to the actual index weights. Panel B of Table 5 shows the same exercise for the Russell 2000. Switching from an equally-weighted to a value-weighted SMB in column 3 increases the estimated alpha by a full percentage point per year, from -2.66% to -1.62%. However, even in the more flexible models, the negative alpha of the Russell 2000 remains significant. As we show later, the remaining alpha is concentrated in June and July, suggesting that it is related to the annual reconstitution of the Russell indices at the end of June.

Table 6 presents an overview of the results for the nine indices. The absolute value of average index alphas and the sum of their squares clearly decline, moving from left to right, and the methodological gap between the academic model and portfolio and index construction in the financial industry narrows. The fit between the models' implied loadings on the 3×4 portfolios and the actual holdings also improves. Again, the largest

[16] In an earlier version of the paper, we made these three changes (splitting SMB into SMM and MMB, adding MidHML, and including Medium with High BM stocks) successively, but since doing so provided no additional insight, we combine them in this version.

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Share codes in market factor	CRSPVW	10/11	10/11	10/11	10/11	10/11				
SMB weighting	EW	EW	VW	VW	VW	VW				
SMB stocks included	As in FF	As in FF	As in FF	All	All	All				
Cutoff for Big stocks	50th pct	50th pct	50th pct	50th pct	50th pct	80th pct				
Size deciles included in BHML	N/A	N/A	N/A	N/A	Top 5	Top 2				
Size deciles included in SHML	N/A	N/A	N/A	N/A	Btm 5	Btm 5				
BM deciles included in H	Top 3	Top 3	Top 3	Top 3	Top 3	Top 7				
Panel A: Alphas										
S&P 500	0.82	0.59	0.33	0.32	0.11	0.21	0.07	0.16	0.35	0.04
S&P 500 Growth	1.82	1.58	1.25	1.23	−0.01	−0.11	−0.13	−0.64	−0.53	−0.60
S&P 500 Value	−0.35	−0.58	−0.76	−0.76	0.07	0.37	0.12	0.49	1.05	0.42
Russell 2000	−2.41	−2.66	−1.62	−1.53	−1.50	−1.61	−2.13	−2.17	−1.07	−1.23
Russell 2000 Growth	−3.41	−3.66	−2.51	−2.43	−1.09	−1.13	−1.77	−1.91	−1.34	−1.58
Russell 2000 Value	−1.25	−1.50	−0.63	−0.54	−1.89	−1.80	−2.18	−1.61	−0.71	−0.62
Russell Midcap	0.17	−0.08	0.21	0.24	0.30	0.45	−0.17	−0.09	0.86	0.52
Russell Midcap Growth	1.61	1.56	1.95	1.97	2.79	1.34	0.43	−0.50	0.69	0.27
Russell Midcap Value	−0.52	−0.62	−0.50	−0.48	−0.59	0.02	−0.64	0.09	1.11	0.59
Panel B: Root mean squared alpha										
All 9 indices	1.70	1.79	1.32	1.29	1.30	1.02	1.21	1.15	0.91	0.79

(*Continued*)

Model	(1) Carhart	(2)	(3)	(4) MOD4	(5)	(6) MOD7	Flexible 13-factor	Flexible NNLS	Actual weights	Average weights
Panel C: Sum of squared differences in 3 × 4 portfolio weights										
S&P 500	0.016	0.016	0.015	0.012	0.014	0.005	0.017	0.006		
S&P 500 Growth	0.250	0.247	0.203	0.201	0.122	0.012	0.056	0.002		
S&P 500 Value	0.115	0.121	0.136	0.132	0.115	0.058	0.052	0.053		
Russell 2000	0.079	0.082	0.014	0.018	0.018	0.026	0.044	0.027		
Russell 2000 Growth	0.198	0.198	0.092	0.065	0.059	0.050	0.081	0.043		
Russell 2000 Value	0.130	0.134	0.098	0.129	0.173	0.103	0.095	0.039		
Russell Midcap	0.120	0.124	0.119	0.116	0.114	0.050	0.043	0.035		
Russell Midcap Growth	0.306	0.303	0.350	0.328	0.428	0.162	0.274	0.167		
Russell Midcap Value	0.307	0.319	0.293	0.297	0.311	0.178	0.088	0.064		
All 9 indices avg	0.169	0.172	0.147	0.144	0.150	0.071	0.083	0.048		
Panel D: Gibbons-Ross-Shanken statistics for all 9 indices										
GRS-statistic	2.898	2.794	2.617	2.012	1.602	1.610				
p-value	0.002	0.003	0.005	0.033	0.107	0.105				

Table 6. Alphas and sum-of-squared differences between weights on 3 × 4 portfolios produced by the models and those from the flexible model.

Description: This table summarizes results from for multiple indices. For each model and index reported in Table 5, this table reports the alphas and the sum of the squared differences between the actual average index holdings of the 3 × 4 portfolios and those implied by the model. For subsets of indices, the table also reports the sum-of-squared average alphas (Panel B), the sum of sum-of-squared differences in the 3 × 4 portfolio weights (Panel C), and the Gibbons-Ross-Shanken test statistic and its small-sample p-value. The time period is from 1980 to 2005.

Interpretation: The table points out how index alphas in general get closer to zero as we begin to modify the Carhart model (1). With separate value factors for small caps and large caps, the alphas are no longer jointly statistically significant in (5) and (6).

improvements in alphas come from the first two steps, between (1) and (3), which include switching to a market portfolio with only U.S. stocks as well as to a value-weighted SMB factor. Alphas decline further between (4) and (6) but this requires adding more factors, which may potentially offset the benefit of reduced index alphas. Because we are concerned about overfitting in-sample, later we also conduct an out-of-sample analysis.

Finally, we formally test whether the alphas of the nine benchmark indices in Table 6 are jointly equal to zero in the various models by calculating the Gibbons-Ross-Shanken test statistic and its associated p-value. The p-value is adjusted for the relatively small sample size but it assumes a normal distribution of the residuals, and it gives the probability that pricing errors are larger than observed if the model were to hold exactly. At the 5%-level, the alphas of the nine indices remain jointly significant, starting with Carhart, until becoming insignificant in the modified factor models in (5) and (6).

4.3.2 Alphas of Other Passive Portfolios

As previously mentioned, the nonzero index alphas are not simply a problem associated with the S&P 500 and the Russell 2000. Instead, the index alphas are symptoms of a more general pattern in which, when using the Carhart model, the largest stocks get positive alphas and small-cap stocks get negative alphas. To illustrate this, Table 7 shows the Carhart and Fama-French alphas of decile portfolios based on market capitalization. We construct the size deciles ourselves to include all U.S. stocks each month, including those with missing or negative book equity, using the size decile cutoffs of the previous month from Ken French's website. Another benefit of using these size deciles is that we can now investigate much earlier time periods, going back all the way to 1926, whereas our index data generally do not start until 1980.

The alphas of these passive size decile portfolios are consistent with the pattern in alphas of the indices. They are, again, also robust to whether we use the Fama-French three-factor model or the Carhart four-factor model. Furthermore, compared to the period from 1927 and 2005, most, though not all, alphas are more similar in the period from 1980 to 2005. Portfolios of large-cap stocks tend to have positive alphas; for example size decile 10 has an annualized Carhart alpha of 0.98% ($t = 2.71$) in the period from 1980 to 2005 and of 0.42% ($t = 2.01$) for in the period from 1927 to 2005.

	Alphas by model (1980–2005)				Alphas by model (1927–2005)			
	FF	(1) Carhart	(4) MOD4	(6) MOD7	FF	(1) Carhart	(4) MOD4	(6) MOD7
Size decile portfolios								
10 (Large)	0.98	0.98	0.30	0.22	0.29	0.42	0.02	0.01
	(2.71)	(2.63)	(0.86)	(0.90)	(1.51)	(2.01)	(0.13)	(0.08)
9	0.31	0.16	0.10	0.23	−0.01	0.20	0.20	0.26
	(0.44)	(0.22)	(0.13)	(0.35)	(−0.03)	(0.48)	(0.48)	(0.68)
8	0.15	0.04	0.48	0.03	−0.03	0.19	0.48	0.42
	(0.17)	(0.05)	(0.56)	(0.08)	(−0.06)	(0.37)	(0.95)	(1.68)
7	0.85	0.37	0.81	0.71	−0.54	−1.08	−0.61	−0.69
	(1.13)	(0.46)	(1.02)	(1.67)	(−1.19)	(−2.32)	(−1.38)	(−2.02)
6	−0.69	−0.85	−0.16	−0.39	−0.59	−0.77	−0.08	−0.07
	(−0.89)	(−1.08)	(−0.20)	(−0.67)	(−1.29)	(−1.66)	(−0.17)	(−0.18)
5	−0.90	−0.72	0.21	0.08	−1.28	−1.08	−0.16	−0.08
	(1.15)	(0.88)	(0.29)	(0.12)	(2.90)	(2.23)	(0.41)	(0.20)
4	−0.97	−1.07	0.14	0.40	−1.01	−0.87	0.36	0.41
	(1.37)	(1.40)	(0.21)	(0.62)	(2.12)	(1.61)	(0.85)	(0.98)
3	−0.90	−0.78	0.58	0.74	−1.87	−1.36	0.00	−0.14
	(1.32)	(1.12)	(1.09)	(1.26)	(3.76)	(2.47)	(0.01)	(0.31)
2	−1.56	−1.48	0.17	−0.21	−2.85	−2.06	−0.31	−0.48
	(1.86)	(1.58)	(0.18)	(0.26)	(4.37)	(2.89)	(0.46)	(0.76)
1 (Small)	−0.04	−0.52	1.40	1.04	−1.95	−0.65	1.57	1.17
	(0.03)	(0.32)	(0.83)	(0.68)	(1.77)	(0.51)	(1.24)	(0.98)
GRS-stat for 10 size deciles	3.23	2.28	1.83	1.79	4.08	3.36	0.83	1.63
p-value	0.06%	1.37%	5.47%	6.20%	0.00%	0.03%	59.87%	9.30%

Table 7. Alphas of size decile portfolios.

Description: This table reports the annualized alphas (in percentages) using various models for size decile portfolios. The size decile portfolios are value-weighted and formed based on market capitalization cutoffs from Kenneth Frenchs website. FF is the Fama-French three-factor model, and Carhart is the four-factor model. Models (4) and (6) are defined as in Table 5. Absolute values of t-statistics based on robust standard errors are in parentheses. The p-value gives the probability that pricing errors would be this large if the model held exactly. Results are shown for two samples: 1980 to 2005 and 1927 to 2005.

Interpretation: The table shows how the FF and Carhart models produce significant positive alphas for large caps and negative alphas for small caps in general, not just for specific indices. In contrast, the modified factor models produce jointly insignificant alphas.

Small-cap deciles tend to have negative alphas, with a greater statistical significance for the three-factor model and the period from 1927 to 2005.

Just like for the indices, these nonzero alphas largely disappear when we slightly modify the factors. We show the results using models (4)

and (6) from Table 5. Both modified factor models, the four-factor and the seven-factor version, significantly reduce the nonzero alphas and render all of them statistically insignificant in both time periods. We also formally test whether the alphas are jointly equal to zero for all 10 size deciles by calculating the Gibbons-Ross-Shanken test statistic and its associated p-value. Using a 5% p-value as the threshold, we find that both modified factor models can explain the size decile portfolios, whereas the Fama-French and Carhart models are rejected.

4.4 Index Reconstitution

Index reconstitution effects present another possible explanation for the negative alpha of the small-cap indices (Petajisto, 2011). Additions to and deletions from the Russell indices are determined once per year based on closing market capitalizations on May 31 and are implemented at the end of June.[17] Stocks that are added to the Russell 2000 outperform those that are deleted in June due to the anticipation of large index fund trading at the end of the month, and some of the excess returns revert in July. These patterns should depress the returns of the Russell 2000 compared to non-Russell 2000 stocks, and may contribute to the negative alpha we find.

We expect these rebalancing effects to be concentrated in June and July, and therefore perform a simple test to determine whether the index reconstitution effect is an important source of the negative alpha of the Russell 2000. We test this by comparing the June and July alphas with those from other months. In Table 8, we estimate three models for the Russell 2000 and its growth component: the Carhart model, model (4) (the Carhart model with a market factor that includes only U.S. common stocks and a value-weighted SMB factor that includes the No BM portfolios), and model (6) (model 4, with SMB split into Mid-minus-Big (MMB) and Small-minus-Mid (SMM), HML replaced by BHML, MidHML, and SHML, and the Medium-BM stocks included with the High-BM stocks in the HML factor). We add to each model an indicator variable for June and July The constant in the model captures the average alpha from August to May, while the June–July coefficient captures any extra alpha in these two months, which could occur due to reconstitution.

[17] Over most of our sample period, the Russell reconstitution took place at the close of the last trading day in June. In 2004, Russell changed this to the Friday that falls between June 21 and June 27.

	Russell 2000			Russell 2000 Growth		
Model	(1) Carhart	(4) MOD4	(6) MOD7	(1) Carhart	(4) MOD4	(6) MOD7
Constant	−0.106	−0.058	−0.064	−0.133	−0.080	−0.025
	(1.65)	(1.07)	(1.24)	(1.84)	(1.32)	(0.45)
June–July dummy	−0.582	−0.422	−0.395	−0.923	−0.748	−0.515
	(3.86)	(3.52)	(3.46)	(4.84)	(4.75)	(4.05)
Total alpha per year	−2.432	−1.542	−1.559	−3.440	−2.450	−1.331

Table 8. Russell 2000 alphas in June and July.

Description: In this table regression models labeled Carhart, (4), and (6) from Table 5 are run including an indicator variable for June and July. Only the constant and June-July coefficients are reported; the other coefficients are very similar to those reported earlier (and a similar table for Russell 2000 Growth). Absolute values of t-statistics from robust standard errors are in parentheses. The time period is from 1980 to 2005.

Interpretation: The table shows how a large fraction of the negative Russell 2000 alpha occurs in June and July, suggesting that the annual index reconstitution plays a role as well.

We find that the alphas for June and July are negative and significant, and collectively explain at least half of the negative alphas for these indices. The proportion that is not explained by the June–July coefficient drops by approximately one half from model (1) to model (6). For models (4) and (6), the August-to-May alpha is no longer statistically significant, even at the 10% level, but the June–July coefficient remains highly significant. In unreported versions of these regressions that include an indicator variable for each month, the June and July coefficients are both significant and are of roughly equal size. The other months with nonzero alphas are December (positive) and January (negative), which is consistent with the well-known January effect.

5 Choosing an Alternative Approach

What do we propose as better factor models that are not subject to the aforementioned issues? We test two different approaches. The first approach is to modify the original FFC factors as discussed above. In our four-factor version (MOD4, or column 4 in Table 5), we restrict the market portfolio to U.S.

stocks, value-weight the SMB factor, and leave the HML and UMD unmodified.[18] We also experiment with a seven-factor version of our modified FFC model (MOD7, or column 6 in Table 5), which introduces a factor for the relative performance of midcaps, splits the value factor in three for the different size groups, and modifies the Small/Big and High/Low-BM cutoffs.

Our second approach is to create size and value factors based on the most commonly used benchmark indices. Our four-factor version (IDX4) replaces SMB with the return differences between the Russell 2000 and S&P 500 and and HML with the differences between the Russell 3000 Value and Growth indices. The seven-factor version (IDX7) adds a factor for the relative performance of the Russell Midcap index and separate Value-Growth factors for the most benchmarked indices in each capitalization group (S&P 500, Russell Midcap, Russell 2000).

Our purpose here is relatively narrow: to find a benchmark model that controls for market, size, and value factors and improves upon the Fama-French-Carhart models in various performance evaluation applications. This means not generating significant alphas for large segments of the market and also better explaining the time-series and cross-sectional variation in returns in real-world portfolios. Because we are now comparing the performance of academic and index-based models, we switch from using indices as test assets to using portfolios of actively managed mutual funds.

5.1 Explaining Common Variation in Mutual Fund Returns

5.1.1 Methodology

A factor model should capture a significant amount of the time-series variation in portfolio returns. This is not only a necessary condition in Arbitrage Pricing Theory for a factor to be priced, but it is also useful for benchmarking purposes because a benchmark that more closely tracks a portfolio return over time produces tighter standard errors for alpha. As our measure of a model's explanatory power, we use the time-series standard deviation of the difference between the return on a portfolio and the return on its benchmark (as determined by the model), commonly called tracking error

[18] Value-weighting the HML factor essentially turns it into a BHML factor, which enhances the model's ability to explain the performance of Large-Growth and Large-Value indices and portfolios, but at the cost of its ability to explain the performance of small-cap portfolios.

volatility (or just "tracking error" for simplicity). In-sample tracking error could be computed as the root mean squared error of a single time-series regression. However, we specifically want to compute out-of-sample tracking error, so we estimate the benchmark portfolio using only past data, and then test how well that benchmark tracks the portfolio in the subsequent period. Tracking error is, of course, just a scaled version of R^2, and thus our tests are analogous to those of Fama and French (1993). However, tracking error conveniently indicates the standard deviation of a money manager's realized alpha, and it also allows our tests to be run out-of-sample, which penalizes a model for overfitting the data and therefore does not bias the results in favor of models with a large number of factors.

Our test assets are actively managed U.S. all-equity mutual funds. This sample not only represents a large cross-section of portfolios, varying from small-cap to large-cap and from value to growth stocks, but it also includes the kind of actual investment portfolios encountered in practical applications.

We analyze two different measures of return. One is the net excess return on a fund (after fees) relative to the risk-free rate. The other is the benchmark-adjusted return on a fund, which means the net return (after fees) in excess of a fund's benchmark index. Rather than relying on the self-reported benchmarks shown in Table 1, each time a fund reports holdings, we follow the methodology of Cremers and Petajisto (2009) and select the index that produces the lowest Active Share, i.e., the index that has the greatest overlap with the fund's portfolio holdings. The rationale behind the benchmark-adjustment is simple: if the benchmark index already captures most of the style differences across funds, then we may not even need an extensive model to account for the residual style differences.

To estimate tracking error for each model, we first need to estimate betas of funds with respect to each model. We estimate betas based on 12 months of daily data on fund returns and index returns (see Appendix A for more discussion of beta estimation). We repeat the beta estimation each time a fund reports its portfolio holdings in the Thomson database, which usually occurs quarterly or semiannually, using the 12 months prior to the report date. Tracking error is then computed for each fund using monthly out-of-sample returns.

We focus on the time period 1996–2005. If we were to start the period earlier, we would have to include years when some indices had not been officially launched and were not known to investors, which probably had an

impact on fund manager behavior. The Securities and Exchange Commission (SEC) began requiring all mutual funds to disclose a benchmark index in their prospectuses in 1998, so this period coincides with one in which managers probably became more benchmark-aware.

5.1.2 Results

Panel A of Table 9 shows the equal-weighted annualized tracking error across all of our benchmark models using excess return or benchmark-adjusted return as the dependent variable. In terms of excess returns, the average fund experienced volatility of 17.35% per year. Controlling for the market portfolio reduces it by about a half to 8.28%, and the Fama-French three-factor model reduces it further to 6.50% per year. Adding the Carhart momentum factor makes little difference for tracking error. The modified versions of FFC produce lower tracking error than the unmodified version, particularly when the extra factors are added.[19]

The index models produce even lower tracking error. Our four-factor index model (IDX4) yields a tracking error of 6.15%, or 5% less than Carhart (6.40%). Adding a midcap index and midcap and small-cap value factors to the model further reduces tracking error to 5.80%. This is 64 bp, or 10%, lower than the tracking error of the Carhart model, indicating an economically meaningful improvement in (out-of-sample) tracking error when using the seven-factor index model.

First subtracting the funds' holdings-based benchmark index from returns further reduces tracking error in all versions of the model. Even without a factor model, simply subtracting the benchmark index return from fund returns reduces tracking error to 6.91%, which is more than three-quarters of the improvement from the CAPM to the unmodified Fama-French model. This suggests there is some merit in the industry's simple approach of benchmark-adjusting returns (although the industry often uses self-reported benchmarks instead of the benchmark derived from holdings that we are using here). Index models still produce lower tracking errors than the analogous modified FFC models, even after benchmark-adjusting returns, although the differences are smaller in percentage terms. The benefit of

[19] In unreported results, we find that most of the gains from switching to seven factors come from splitting the value factor into large- and small-cap versions.

	Tracking error volatility (% per year)							
Model	None	CAPM	FF	Carhart	MOD4	MOD7	IDX4	IDX7
Panel A: All funds								
Excess return	17.35	8.28	6.50	6.44	6.40	6.15	6.15	5.80
Benchmark-adjusted	6.91	6.58	6.18	6.14	6.12	5.99	6.03	5.71
Panel B: Active Share $<$ median								
Excess return	16.02	6.53	5.19	5.20	5.20	4.95	4.73	4.49
Benchmark-adjusted	5.33	5.13	4.78	4.75	4.73	4.61	4.62	4.36
Panel C: All funds, alternative estimation periods								
Daily data, 6 months								
Excess return	17.35	8.23	6.49	6.48	6.49	6.21	6.19	5.85
Benchmark-adjusted	6.91	6.55	6.18	6.21	6.18	6.08	6.02	5.75
Monthly data, 3 years								
Excess return	17.35	8.77	6.82	6.82	6.76	6.94	6.48	6.57
Benchmark-adjusted	6.91	6.87	6.74	6.78	6.76	7.18	6.80	7.05
Monthly data, 5 years								
Excess return	17.35	8.64	6.90	6.86	6.75	6.77	6.42	6.45
Benchmark-adjusted	6.91	6.83	6.74	6.75	6.71	6.96	6.71	6.85

None	—	MOD4	MKT2, SMB2, HML, UMD
CAPM	MKT	MOD7	MKT2, MMB, SMM, BHML, MHML, SHML, UMD
FF	MKT, SMB, HML	IDX4	S5, R2-S5, R3V-R3G, UMD
Carhart	MKT, SMB, HML, UMD	IDX7	S5, RM-S5, R2-RM, S5V-S5G, RMV-RMG, R2V-R2G, UMD

Table 9. Mutual fund out-of-sample tracking error across benchmark models.

Description: This table shows the out-of-sample tracking error volatility for U.S. all-equity mutual funds for 1996-2005. Whenever a fund reports its positions (semiannually or quarterly), its prior 12-month daily returns are regressed on each of the factor models to determine its betas. Using those betas, the funds monthly out-of-sample predicted return and the difference between the predicted and actual fund return are computed. Each funds tracking error is computed as the time-series volatility of that difference over the sample period. Each number in the table represents an equal-weighted average of those tracking errors across funds. Fund returns are expressed both as excess returns and benchmark-adjusted returns. Panel B uses only funds with low Active Share. Panel C shows the results for different lengths and sampling intervals of the estimation period. MKT, SMB, and HML are the standard Fama-French factors, while MKT2 and SMB2 are our modified versions of the FF factors. R2-S5 is an index-based factor of Russell 2000 minus S&P 500, R3V is Russell 3000 Value, and RMG is Russell Midcap Growth.

Interpretation: The table points out that the index models IDX4 and IDX7 produce lower out-of-sample tracking error than the Carhart model, indicating that they more closely track actual investment portfolios.

using additional factors and index models rather than modified FFC is slightly smaller after benchmark adjusting.

Panel B repeats the same exercise but uses only relatively passive funds, which are easiest to explain with factor models. We compute each fund's Active Share following Cremers and Petajisto (2009), and we select funds in the bottom 50% of Active Share within each benchmark index. We find that all tracking errors go down by about 120–140 bp per year but that conclusions about the relative performance of models remain unchanged. Panel C shows results from a shorter estimation window and uses monthly data.[20] Index models continue to outperform modified FFC, but in the monthly data, four-factor models perform better than seven-factor models. It is likely that this is due to the fact that beta estimates are less precise in the monthly data, differentially degrading out-of-sample performance in models with more factors.

5.2 Explaining the Cross-Section of Mutual Fund Returns

The nonzero index alphas that we document should matter most for performance evaluations when comparisons are made between managers investing in different size or value categories. In this section we examine how the choice of a benchmark model affects conclusions about the skill of the average mutual fund manager in different style categories.

5.2.1 Methodology

To form groups among similar funds, and to maximize cross-sectional differences across groups, we create nine portfolios of funds from a two-dimensional sort on size and value. In particular, we determine the fund groups from their holdings-inferred benchmark indices. The large-cap group consists of funds that use the S&P 500, Russell 1000, Russell 3000, or Wilshire 5000 as their benchmarks. The midcap indices are the S&P 400, Russell Midcap, and Wilshire 4500. The small-cap indices are S&P 600 and Russell 2000. The value and growth groups are determined from the corresponding style indices. We use net fund returns (after fees and transaction costs), as is common in the literature, though results would be similar (with

[20] See the Online Appendix for further description of these results.

small funds gaining a slight edge over large funds) if expenses were added back to the net fund returns.

We again examine both excess returns and benchmark-adjusted returns. First, the benchmark-adjusted return is the performance measure that most mutual fund clients focus on because their natural investment alternative is a low-cost index fund that replicates the index return. It is also the measure that fund managers focus on because beating the index is their explicit self-declared investment objective. Second, if a benchmark model gives very different results for excess returns and benchmark-adjusted returns, it can only come from nonzero alphas assigned to the benchmark indices themselves. Because we want to avoid attributing any skill to the passive benchmark index, a good benchmark model should produce similar alphas for both excess returns and benchmark-adjusted returns.

5.2.2 Results

Table 10 shows the fund alphas across our key benchmark models. The time period is from 1996 to 2005 so that all indices are available for the entire sample. Each fund group represents an equal-weighted portfolio of funds. We estimate betas and alphas from monthly returns on these portfolios of funds and the benchmark factors. Fund returns are net returns, i.e., after all fees and expenses are deducted. We use both excess returns and benchmark-adjusted returns on funds and show alphas of the funds in the 3×3 sort on size and value.

If we look at excess returns without any risk adjustments, as opposed to the risk-adjusted alphas reported in the table, we find that small-cap funds beat large-cap funds by 2.79% per year and value funds beat growth funds by 1.90% per year over this 10-year period. Controlling for the benchmark index returns, the average fund lost to its benchmark by 0.80% per year, which is slightly less than the average expenses charged by the funds. Furthermore, the benchmark adjustment completely eliminates the return spread between growth and value funds, pushing it from 1.90% to -0.14%, and it reduces the return spread between small-cap and large-cap funds from 2.79% to 2.02%.

The most interesting patterns in alphas occur for the Carhart model (Panel A). With excess returns, the Carhart model shows that alphas of small-cap funds are 2.13% ($t = 1.88$) *below* the large-cap fund alphas,

	Excess return						Benchmark-adjusted return				
Size	Value group					Size	Value group				
group	1	2	3	All	High-Low	group	1	2	3	All	High-Low
Panel A: Carhart (MKT, SMB, HML, UMD)											
3	−1.24	−1.01	−1.30	−1.07	−0.06	3	−3.28	−1.81	−0.92	−2.26	2.36
	(−1.51)	(−2.33)	(−1.26)	(−2.06)	(−0.04)		(−3.31)	(−4.42)	(−1.31)	(−3.99)	(1.83)
2	−2.37	−1.69	−0.53	−1.69	1.84	2	−3.35	−2.22	−0.31	−2.58	3.04
	(−1.24)	(−1.23)	(−0.37)	(−1.14)	(0.85)		(−3.03)	(−2.46)	(−0.34)	(−3.18)	(2.32)
1	−3.99	−3.09	−1.20	−3.20	2.79	1	1.69	−0.73	1.06	0.68	−0.63
	(−2.06)	(−2.15)	(−0.91)	(−2.26)	(1.39)		(1.39)	(−0.79)	(0.91)	(0.82)	(−0.42)
All	−2.08	−1.75	−1.27	−1.69	0.81	All	−2.34	−1.60	−0.43	−1.70	1.91
	(−1.83)	(−2.56)	(−1.20)	(−2.21)	(0.53)		(−3.09)	(−4.47)	(−0.70)	(−3.68)	(1.95)
High-Low	2.75	2.08	−0.10	2.13		High-Low	−4.97	−1.09	−1.98	−2.94	
	(1.70)	(1.64)	(−0.10)	(1.88)			(−3.24)	(−1.08)	(−1.56)	(−3.07)	
Panel B: MOD4 (MKT2, SMB2, HML, UMD)											
3	−1.21	−1.10	−1.34	−1.12	−0.14	3	−2.98	−1.65	−0.74	−2.04	2.24
	(−1.44)	(−2.73)	(−1.35)	(−2.19)	(−0.10)		(−3.23)	(−4.11)	(−1.05)	(−3.82)	(1.79)
2	−1.58	−0.98	−0.16	−0.99	1.43	2	−3.27	−2.01	−0.12	−2.45	3.15
	(−0.84)	(−0.75)	(−0.11)	(−0.69)	(0.66)		(−3.02)	(−2.29)	(−0.13)	(−3.12)	(2.39)
1	−2.74	−2.03	−0.19	−2.10	2.54	1	1.55	−0.84	0.96	0.55	−0.59
	(−1.57)	(−1.49)	(−0.15)	(−1.62)	(1.31)		(1.27)	(−0.94)	(0.85)	(0.67)	(−0.39)
All	−1.61	−1.45	−1.05	−1.35	0.56	All	−2.18	−1.50	−0.31	−1.57	1.87
	(−1.41)	(−2.19)	(−1.05)	(−1.82)	(0.37)		(−3.07)	(−4.37)	(−0.51)	(−3.63)	(1.96)
High-Low	1.53	0.94	−1.15	0.98		High-Low	−4.53	−0.81	−1.71	−2.59	
	(1.15)	(0.78)	(−0.97)	(0.97)			(−3.05)	(−0.81)	(−1.35)	(−2.78)	
Panel C: MOD7 (MKT2, MMB, SMM, BHML, MHML, SHML, UMD)											
3	−1.48	−1.16	−0.47	−1.00	1.01	3	−1.77	−1.31	−1.05	−1.47	0.72
	(−1.95)	(−2.88)	(−0.64)	(−2.06)	(0.94)		(−2.37)	(−4.35)	(−1.60)	(−3.75)	(0.63)
2	−0.45	−0.64	0.19	−0.08	0.64	2	−2.69	−1.89	−0.65	−2.15	2.03
	(−0.31)	(−0.56)	(0.15)	(−0.07)	(0.36)		(−2.43)	(−2.24)	(−0.73)	(−2.83)	(1.56)
1	−1.54	−2.53	−1.39	−2.05	0.15	1	0.81	−0.16	1.16	0.54	0.35
	(−1.00)	(−1.82)	(−1.13)	(−1.62)	(0.09)		(0.63)	(−0.18)	(0.95)	(0.64)	(0.22)
All	−1.39	−1.57	−0.66	−1.17	0.73	All	−1.52	−1.11	−0.54	−1.18	0.98
	(−1.44)	(−2.33)	(−0.80)	(−1.69)	(0.61)		(−2.46)	(−3.48)	(−0.86)	(−3.15)	(1.08)
High-Low	0.06	1.36	0.92	1.05		High-Low	−2.58	−1.16	−2.22	−2.01	
	(0.05)	(1.10)	(0.90)	(1.03)			(−1.75)	(−1.19)	(−1.79)	(−2.22)	

Table 10. Mutual fund alphas: Carhart and modified factor models.

Description: This table shows the alphas of net return for U.S. all-equity mutual funds 1996–2005. Funds are sorted into groups based on their estimated benchmark indices: the size groups represent small, mid, and large-cap stocks, and the value groups represent growth, core, and value stocks. Alphas are computed with excess return or benchmark-adjusted return as left-hand-side variables and various benchmark models on the right-hand side. The numbers show the annualized alpha, with t-statistics in parentheses below. Panels A, B, and C show the models labeled Carhart, M0D4, and M0D7 in Table 9, respectively.

Interpretation: The table points out that Carhart alphas vary significantly across mutual fund styles and are not robust to a simple subtraction of benchmark index return. The modified factor models are slightly more stable.

but with benchmark-adjusted returns, the small-cap fund alphas are 2.94% ($t = 3.07$) *above* the large-cap fund alphas. The simple benchmark adjustment therefore changes the small- and large-cap alphas by 5.07% for the Carhart model. This is a truly dramatic effect, especially in the context of mutual fund alphas, which are, on average, very close to zero, and it is certainly large enough to potentially reverse the conclusions of performance analysis. The results are similar for the Fama-French model, and they can only come from nonzero alphas that the two models assign to the benchmark indices. We argue that this finding casts doubt on the validity of the standard Fama-French-Carhart alpha estimates when comparing across the size dimension. Across the value dimension, effects vary in the three size categories.

Panels B and C in Table 10 present the results for the modified Fama-French market, size, and value factors. For excess returns in Panel B, we find that after the small modifications to the factors, large-cap funds no longer significantly outperform small-cap stocks (the difference in alpha across these groups equals 0.98% with $t = 0.97$). However, these modifications have little impact on the alphas of benchmark-adjusted returns since the benchmark index already captures most of the systematic risk exposures of funds.

Adding a midcap factor and separate value factors for small-, mid-, and large-cap stocks in Panel C results in alphas for excess returns that are comparable to those in Panel B, except that the alphas for benchmark-adjusted returns are now more in line with the alphas of the excess returns in the value-growth direction (that is insignificant in both cases). The main difference remains that the alphas of benchmark-adjusted returns suggest that small-cap funds strongly outperform large-cap funds (a difference in alpha of -2.01% with $t = 2.22$) whereas excess returns show the opposite (an alpha difference of 1.05% albeit with only $t = 1.03$).

Table 11 and its panels B and C report the corresponding alphas from the index models. In contrast to the Carhart model, the fund alphas are now very similar across excess returns and benchmark-adjusted returns, especially with the seven-factor model in Panel C. This arises from the fact that the index models produce exactly zero alphas for the constituent indices (by construction) and only small alphas for the other indices. Like in the tracking error analysis, this has the important implication that the seven-factor index model can be applied to the excess returns on all fund returns,

Size group	Excess return: Value group 1	2	3	All	High-Low	Size group	Benchmark-adjusted return: Value group 1	2	3	All	High-Low
Panel A: Carhart (MKT, SMB, HML, UMD)											
3	−1.24	−1.01	−1.30	−1.07	−0.06	3	−3.28	−1.81	−0.92	−2.26	2.36
	(−1.51)	(−2.33)	(−1.26)	(−2.06)	(−0.04)		(−3.31)	(−4.42)	(−1.31)	(−3.99)	(1.83)
2	−2.37	−1.69	−0.53	−1.69	1.84	2	−3.35	−2.22	−0.31	−2.58	3.04
	(−1.24)	(−1.23)	(−0.37)	(−1.14)	(0.85)		(−3.03)	(2.46)	(−0.34)	(−3.18)	(2.32)
1	−3.99	−3.09	−1.20	−3.20	2.79	1	1.69	−0.73	1.06	0.68	−0.63
	(−2.06)	(−2.15)	(−0.91)	(−2.26)	(1.39)		(1.39)	(−0.79)	(0.91)	(0.82)	(−0.42)
All	−2.08	−1.75	−1.27	−1.69	0.81	All	−2.34	−1.60	−0.43	−1.70	1.91
	(−1.83)	(−2.56)	(−1.20)	(−2.21)	(0.53)		(−3.09)	(−4.47)	(−0.70)	(−3.68)	(1.95)
High-Low	2.75	2.08	−0.10	2.13		High-Low	−4.97	−1.09	−1.98	−2.94	
	(1.70)	(1.64)	(−0.10)	(1.88)			(−3.24)	(−1.08)	(−1.56)	(−3.07)	
Panel B: IDX4 (S5, R2-S5, R3V-R3G, UMD)											
3	−1.52	−1.22	−0.51	−1.10	1.01	3	−2.02	−1.23	−0.22	−1.38	1.81
	(−2.05)	(−3.57)	(−0.87)	(−2.32)	(1.28)		(−2.63)	(−3.69)	(−0.35)	(−3.22)	(1.63)
2	−0.64	0.92	1.90	0.36	2.55	2	−2.82	−1.67	−0.06	−2.09	2.75
	(−0.39)	(0.81)	(1.67)	(0.28)	(1.36)		(−2.64)	(−1.97)	(−0.07)	(−2.63)	(2.25)
1	−0.74	0.75	3.24	0.46	3.98	1	1.55	−0.95	0.74	0.44	−0.80
	(−0.49)	(0.71)	(2.68)	(0.48)	(1.93)		(1.31)	(−1.11)	(0.68)	(0.58)	(−0.52)
All	−1.15	−0.58	0.45	−0.54	1.59	All	−1.54	−1.21	0.01	−1.13	1.55
	(−1.18)	(−1.13)	(0.68)	(−0.88)	(1.47)		(−2.41)	(−3.77)	(0.01)	(−2.85)	(1.87)
High-Low	−0.78	−1.97	−3.76	−1.56		High-Low	−3.57	−0.27	−0.96	−1.82	
	(−0.62)	(−1.94)	(−3.15)	(−1.87)			(−2.75)	(−0.30)	(−0.79)	(−2.34)	
Panel C: IDX7 (S5, RM-S5, R2-RM, S5V-S5G, RMV-RMG, R2V-R2G, UMD)											
3	−1.93	−1.45	−0.66	−1.29	1.27	3	−1.71	−1.44	−0.84	−1.44	0.87
	(−3.32)	(−5.44)	(−1.16)	(−3.60)	(1.55)		(−3.40)	(−5.46)	(−1.64)	(−4.40)	(1.22)
2	−1.40	−0.14	0.70	−0.41	2.11	2	−1.44	−0.67	0.66	−0.90	2.11
	(−1.61)	(−0.16)	(0.78)	(−0.55)	(1.89)		(−1.73)	(−0.71)	(0.76)	(−1.45)	(1.89)
1	0.29	0.12	1.64	0.37	1.35	1	0.29	−0.38	1.64	0.32	1.35
	(0.23)	(0.11)	(1.53)	(0.39)	(0.92)		(0.24)	(−0.45)	(1.59)	(0.41)	(0.92)
All	−1.51	−1.07	−0.11	−0.88	1.39	All	−1.30	−1.07	−0.16	−0.99	1.15
	(−2.46)	(−2.26)	(−0.20)	(−1.94)	(1.78)		(−2.17)	(−3.41)	(−0.30)	(−2.26)	(1.47)
High-Low	−2.22	−1.57	−2.29	−1.66		High-Low	−2.00	−1.06	−2.48	−1.76	
	(−1.89)	(−1.51)	(−2.06)	(−1.93)			(−1.79)	(−1.19)	(−2.34)	(−2.32)	

Table 11. Mutual fund alphas: Carhart and index models.

Description: This table shows the alphas of net return for U.S. all-equity mutual funds 1996-2005. Funds are sorted into groups based on their estimated benchmark indices: the size groups represent small, mid, and large-cap stocks, and the value groups represent growth, core, and value stocks. Alphas are computed with excess return or benchmark-adjusted return as left-hand-side variables and various benchmark models on the right-hand side. The numbers show the annualized alpha, with t-statistics in parentheses below. Panels A, B, and C show the models labeled Carhart, IDX4, and IDX7 in Table 9, respectively.

Interpretation: The table points out that while the Carhart alphas vary significantly across mutual fund styles and are not robust to a simple subtraction of the benchmark index return, the index models produce less extreme alphas throughout.

regardless of a fund's style or benchmark index. Furthermore, the finding that the seven-factor index model produces alphas that are surprisingly similar to the benchmark-adjusted returns suggests that even the simple subtraction of the benchmark index return may be a better benchmark model than the standard academic three- or four-factor models.

In terms of the magnitude of alphas, the seven-factor index model produces values that seem relatively plausible *ex-ante*. The average fund underperformed by −0.88% using excess returns ($t = 1.94$), large-cap funds underperformed by −1.29% ($t = 3.60$), and small-cap funds actually slightly outperformed by 0.37%, although without statistical significance. There is no statistically strong pattern across value groups. The four-factor index model also produces alpha estimates that are relatively similar across size and value groups, which is intuitively consistent with a competitive equilibrium between actively managed mutual funds. (Note here the contrast to the Carhart model, which shows the average alpha across all small-fund funds being −3.20% per year.) In general, the four-factor and seven-factor index models seem to perform the best among the models tested here.

Finally, a possible concern about the index-based factors is that they include index reconstitution effects, which reduces the performance of small-stock indices like the Russell 2000 and allows a small-cap manager to earn an alpha by simply avoiding those effects. This could partly explain our finding that small-cap managers have outperformed large-cap managers in terms of benchmark-adjusted returns.

6 Conclusions

The standard Fama-French and Carhart models, which have been widely adopted in academic research for asset pricing and performance evaluation purposes, suffer from biases. The SMB factor assigns disproportionate weight to value stocks, especially within large stocks, which in turn induces a positive correlation in the SMB and HML betas of cap-weighted portfolios. Likewise, the HML factor assigns disproportionate weight to small-cap stocks, which increases its returns due to the outperformance of small-cap value since 1980. Taken together, these two effects cause the benchmarks that are provided by the Fama-French or Carhart models to be tough to beat for small-cap managers (who have a positive beta on SMB) and easy to

beat for large-cap managers (who have a negative beta on SMB). Furthermore, the CRSP value-weighted market index, which includes other securities besides U.S. stocks, contributes to a positive bias to all alpha estimates for U.S. stocks.

One of the most striking pieces of evidence for this bias comes from the four-factor Carhart alphas of passive benchmark indices. The most common large-cap indices, the S&P 500 and Russell 1000, have exhibited economically and statistically significant positive alphas of 0.82% and 0.47% per year from 1980 to 2005. The corresponding small-cap indices, the Russell 2000 and S&P 600, have earned significant negative alphas of −2.41% and −2.59% per year. Naturally, one would expect passive benchmark indices to have zero alphas; in fact, one could even define alpha relative to a set of passive indices, which are the low-cost alternatives to active management.

As alternatives to the well-known three- and four-factor models, we test models with modified versions of the Fama-French factors as well as models based on the common benchmark indices. We analyze tracking error volatility across a broad cross-section of mutual funds to see which models best explain the common variation in returns and thus most closely track the time series of fund returns. The index-based models produce the lowest out-of-sample tracking error volatility, and therefore outperform the traditional Fama-French and Carhart models.

When applied to the cross-section of average mutual fund returns, the index-based models explain average returns well and produce alphas close to zero for all fund groups. The Carhart model produces slightly larger negative alphas in general, but its biggest weakness is its sensitivity to a seemingly innocuous adjustment: when comparing small-cap and large-cap funds, adjusting for the benchmark index has a drastic 5% per year impact on their Carhart and Fama-French alphas, reversing the conclusions about how average manager skill differs between small- and large-cap funds. The index-based models do not exhibit similar sensitivity, as they do not produce significant nonzero alphas for large-cap stocks and small-cap stocks in general.

Overall, the results support the use of alternative models for performance evaluation. When betas can be estimated using daily data, mutual fund returns are best explained by our seven-factor index model, which includes the S&P 500, Russell Midcap, and Russell 2000, separate

value-minus-growth factors for each index, and a momentum factor (UMD). When only monthly data is available, economizing on the number of factors may improve performance, but an index-based four-factor model with the S&P 500, Russell 2000, Russell 3000 Value-minus-Growth, and UMD still outperforms the Carhart model.

To researchers who are reluctant to use index-based factors due to concerns about basing academic research on proprietary indices which may involve subjective judgments in their construction, we point out that our index models do perform better empirically. This should not be a surprise. If co-movements in stocks within a given size or value category are partly produced by changes in investors' appetites for stocks of a given style, and if these appetites get expressed via investment vehicles that track benchmark indices, then the indices themselves should track the resulting asset price changes more precisely than academic factors that approximate them (Roll, 1992; Stutzer, 2003). Investors may also prefer to trade common indices due to lower transaction costs, especially when shorting stocks.

This raises the question of whether index-based factors would be more appropriate for asset pricing as well as performance evaluation applications. In earlier versions of this paper, we presented evidence that our index factor models outperformed both modified and unmodified FFC models in explaining the cross-section of stock returns. We have opted to focus this paper entirely on performance evaluation, but the question of whether it is preferable to use either indices or factors modified to be more index-like in pricing models is worthy of investigation.

Despite the improvements we accomplish with our alternative factor models, we do not hold them out as "perfect" alternatives. Instead, our main contribution is to identify the sources of index alphas and the pitfalls of certain methodological choices, thus paving the way for future researchers to build even better models. While awaiting new models, we recommend researchers use either our index-based seven-factor model, or our four-factor model with just the S&P 500, Russell 2000, Russell 3000 Value-minus-Growth, and UMD, or at the very least make the small modifications we recommend to the FFC factors. Given how widely the Fama-French and Carhart models are used to measure abnormal portfolio performance, eliminating the biases we document from future studies seems worth the modest effort required.

References

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: 341–372.

Boyer, B. H. 2006. "Comovement Among Stocks with Similar Book-to-Market Ratios," working paper.

Carhart, M. 1997. "On Persistence in Mutual Fund Returns." *Journal of Finance* 52(1): 57–82.

Chan, L. K. C., S. G. Dimmock, and J. Lakonishok. 2009. "Benchmarking Money Manager Performance: Issues and Evidence." *Review of Financial Studies* 22(11): 4553–4599.

Cremers, K. J. M. and A. Petajisto. 2009. "How Active Is Your Fund Manager? A New Measure That Predicts Performance." *Review of Financial Studies* 22(9): 3329–3365.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers. 1997. "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks." *Journal of Finance* 52(3): 1035–1058.

Elton, E. J., M. J. Gruber, and C. R. Blake. 1999. "Common Factors in Active and Passive Portfolios." *European Finance Review* 3: 53–78.

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.

Ferson, W. E. and R. W. Schadt. 1996. "Measuring Fund Strategy and Performance in Changing Economic Conditions." *Journal of Finance* 51: 425–461.

Goetzmann, W. N., Z. Ivkovic, and K. G. Rouwenhorst. 2001. "Day Trading International Mutual Funds: Evidence and Policy Solutions." *Journal of Financial and Quantitative Analysis* 36(3): 287–309.

Loughran, T. 1997. "Book-to-Market Across Firm Size, Exchange, and Seasonality: Is There an Effect?" *Journal of Financial and Quantitative Analysis* 32: 249–268.

Moor, L. De and P. Sercu. 2006. "The Small Firm Anomaly: US and International Evidence." working paper, Katholieke Universiteit Leuven.

Pastor, L. and R. F. Stambaugh. 2002. "Mutual Fund Performance and Seemingly Unrelated Assets." *Journal of Financial Economics* 63: 315–349.

Petajisto, A. 2011. "The Index Premium and Its Hidden Cost for Index Funds." *Journal of Empirical Finance* 18: 271–288.

Ritter, J. R. 1991, "The Long-run Performance of Initial Public Offerings." *Journal of Finance* 46: 3–28.

Roll, R. 1992. "A Mean-Variance Analysis of Tracking Error." *Journal of Portfolio Management* 18: 13–22.

Ross, S. 1976. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13: 341–360.

Sharpe, W. F. 1992. "Asset Allocation: Management Style and Performance Measurement," *Journal of Portfolio Management* 18(2): 7–19.

Stutzer, M. 2003, "Fund Managers May Cause Their Benchmarks to be Priced Risks." *Journal of Investment Management* 1: 1–13.