Where's the Beef?

Rob Arnott, Amie Ko, and Lillian Wu

Rob Arnott
is a partner and chair at
Research Affiliates, LLC,
in Newport Beach, CA.
arnott@rallc.com

Amie Ko
is senior vice president,
head of marketing, at
Research Affiliates, LLC,
in Newport Beach, CA.
ko@rallc.com

Lillian Wu
is vice president, research,
at Research Affiliates Global
Advisors (Europe) Limited in
London, UK.
wu@rallc.com

KEY FINDINGS

- Performance chasing and data mining are blunders afflicting traditional and quantitative investors of today's so-called smart beta strategies.
- p-hacking, noise trading, fad chasing, and nowcasting are practices harbored within academia and the investment industry that exacerbate our innate tendency to embrace performance chasing and data mining.
- Assessing the impact of revaluation alpha, acknowledging the effect of implementation costs and other hidden costs, and addressing clients' vulnerability to a performance expectations shortfall can lead to better investor outcomes.

ABSTRACT

In the 1980s there was a famous TV ad for Wendy's with the tagline "Where's the beef?"[1] Many investors in today's so-called smart beta strategies may well be asking a similar question, "Where's the alpha?" Investors frequently buy into historical simulations or backtests, often supported by compelling studies by respected academics, suggesting wonderful performance with remarkable consistency, only to earn no alpha once they invest. The only winners typically are the asset managers and brokers through their fees and commissions. The problem is data mining and performance chasing, the nemeses of all investors. Yes, academics, "quants," and investment professionals are all subject to those same temptations, very nearly to the same extent as retail investors. This article explores the ways seasoned professionals fall prey to these simple blunders and suggests the three lessons that could perhaps allow us to better meet client expectations, both by delivering improved outcomes and by encouraging more sensible expectations.

Arguably the two greatest mistakes in investing are performance chasing and data mining. The two are interrelated, and quantitative investors ("quants"), reliant on computer models for their investment decisions, are no less prone to those errors than are traditional investors. We all are familiar with the SEC Rule 156 performance disclaimer that requires some variant of "past performance is no guarantee of future results." And yet, human nature pulls us in the opposite direction. Any newly expensive asset, priced to disappoint in the future, likely got there by providing investors with joy and profit. It is painful to contemplate selling such assets. Reciprocally, any bargain likely got there by inflicting pain and losses. It goes against human nature to say, "I want more of that!"

[1] https://www.youtube.com/watch?v=idnwh6iDnXA.

In the 1980s, Barr Rosenberg, one of the great first-generation quants, was asked what advantage quantitative investors have over seasoned professionals, who carefully analyze the business prospects and relative values of individual companies. He famously quipped, "About 4% a year." This is self-evidently no longer true. Quants now compete against one another, each seeking an edge. Trading is facilitated by high-frequency traders and market makers—and those quants with a short-term focus—all applying quantitative models on intraday tick data. During the past 20 years, we've written repeatedly on topics that—we believe—can help our clients better achieve their goals. In this article, we will touch on a few of those topics.

HOW DID WE GET HERE?

The tendency to chase performance and mine data afflicts novice investors, seasoned professionals, and quants alike, albeit in varying degrees, because evolution has adapted this behavior into *Homo sapiens*. Our ancestors on the African veldt did not fare well if they ran toward a lion. During the past 50 years, a large body of research shows that people tend to form their expectations by relying on the recent past or by using heuristics to inform their investment decision making. The representativeness heuristic (Kahneman and Tversky 1972) and availability bias (Tversky and Kahneman 1973) are classic rule-of-thumb shortcuts that steer investors to favor recent winners and trendy themes and extrapolate recent experiences as if past is prologue.

De Bondt and Thaler (1985) found that this pattern of decision making—that is, investors' preference for past winners—also affects market prices. They found that the biggest losers of the previous 36 months beat the biggest winners by 19.6% in the subsequent 36 months, a result that was (and remains!) both statistically and economically significant. Moreover, these natural tendencies extend beyond individual investors. Empirical studies, including those by Grinblatt, Titman, and Wermers (1995) and Wermers (1999), found trend-chasing behavior prevalent among mutual fund managers, and Badrinath and Wahal (2002) documented this behavior by institutional investors.

While we, and many others, have been warned of these tendencies and biases in the past, the problem has arguably become more acute in recent years. A host of practices and incentives harbored within academia and the investment industry are exacerbating our innate tendency to embrace performance chasing and data mining. Notably, several relevant practices plaguing our industry include p-hacking, noise trading, fad chasing, and nowcasting.

p-Hacking

Data mining is particularly lethal to academics and quants. We all rely on data. Academics use past data to validate research hypotheses for publication. Positive and statistically significant results increase the odds of a paper's acceptance by a journal. The need to publish therefore creates incentives for data mining, leading to the proliferation of implausible backtests with overoptimistic return outcomes.[2] Harvey, Liu, and Zhu (2016) documented more than 300 factors published in the top three academic journals alone and showed that many of the published factors are

[2] Harvey (2017) discussed an agency dilemma related to the publication biases. Because editors prefer to publish papers with the most significant results to compete for citation-based impact factors, authors in response choose to ignore or discard weak results. In a more disconcerting way, some authors likely choose a specific sample and testing method until results turn from insignificant to significant (p-hacking).

lucky findings resulting from the multiple-testing problem. Selection bias guarantees that random positive noise will be overwhelmingly the norm when performance is the basis for selection. As a result, among many factors tested, some will appear to be statistically significant by random luck. Moreover, the contrast between pre- and post-publication outcomes is stark, with the excess return following publication falling far short of in-sample published results. As McLean and Pontiff (2016) reported, the post-publication premiums of 97 equity factor strategies fell by an average of 32% versus the published figure.

Even traditional managers form intuitions and heuristics based on past experience, hence past data. Naturally, we all use backtests to check whether our ideas have any merit. We could think of this as "light data mining." The simpler the model, the more useful the results. Ideally, we want an out-of-sample test (perhaps other countries, other markets, or earlier or later data) as a cross-check to allow us to have more confidence in our results. When we use backtests to improve our models, we take data mining to a new and dangerous level. Cam Harvey has long described this work as p-hacking, seeking a maximum probability (p) that the result is not a consequence of random noise. When we use backtests to improve our backtests, we're engaged in a particularly pernicious variant of data mining.[3] If we believe that we're making our strategies better, we're fools. If we use this kind of backtest to persuade investors to commit their money to our ideas without substantive disclaimers, we're not honest.

Noise Trading

In his seminal 1986 article, "Noise," Fischer Black took our industry to task for imagining that we have knowledge that is unknown to the general marketplace. When we trade according to facts that are already discounted in current share prices, we are trading on "noise." If we are to beat the market, someone on the other side of our trades must, on average, be losing. Do we know something they don't know? Most investors harbor this illusion, that they have insights missed by others. But, their trading counterparty is far more likely to be a massively informed and sophisticated institutional investor, not a naïve retiree (or newbie). Or does our trading counterparty have a utility function that differs from ours, in ways that allow them to tolerate underperformance?[4]

The old poker aphorism applies (adapted for investors by Warren Buffett in his 1985 letter to shareholders): "If after 10 minutes at the poker table you do not know who the patsy is—you are the patsy." *An important counterpoint to this is that skill-based investors, who add value for their clients based on insights and not on luck, likely exist*.[5] Sadly, for every 100 investors who claim a competitive advantage, there

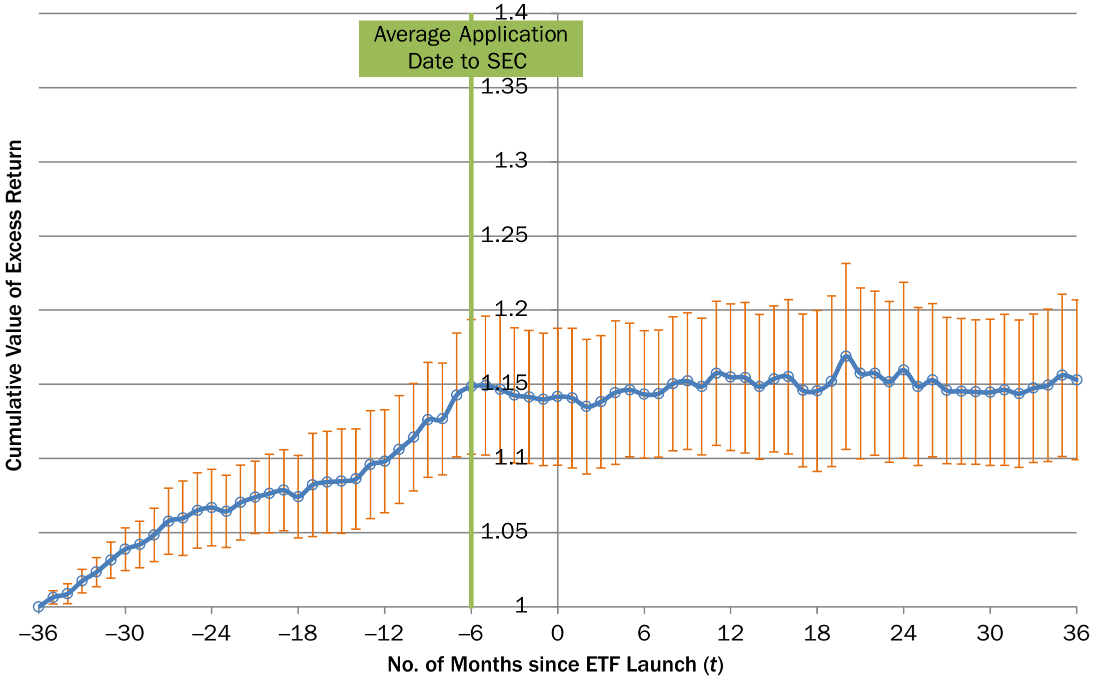
[3] See Arnott, Harvey, and Markowitz (2019).

[4] This is not as far-fetched as it sounds. In bonds, we have central bankers interested in tamping down volatility and lacking a profit motive, and we have insurance companies and banks that face regulatory haircuts for many asset categories. Both will shun profitable investments that are contrary to these overriding motivations. In stocks, we have index funds that will shun performance-enhancing strategies if they lead to tracking error relative to their target index.

[5] Statistical evidence supports that assertion. The dispersion of realized returns for funds and asset managers should shrink roughly with the square root of time. That is, 10-year return dispersion among funds and managers should be less than one-third as wide as the average of the 1-year dispersion of return. Survivorship bias should cause the dispersion to narrow even a bit further, as many of the losers fail to survive the 10 years. The reality is that the 10-year dispersion typically narrows by 50% to 60%, not by the theoretical 70% or so. This outcome is not consistent with a world in which performance results are a random draw; it bears more resemblance to a world in which half of the asset managers have an average alpha of 1% to 2%, and half have a negative alpha of similar magnitude, and in which all managers have a random walk around their average alpha.

EXHIBIT 1

Three-Year Cumulative Relative Index Performance before and after ETF Launch



SOURCE: Brightman, Li, and Liu (2015).

are likely only a handful who actually have an advantage, offset (and funded) by a perhaps-larger number who have negative skill.

We have often invited clients—and other practitioners—to ask two related questions: Who is taking the other side of your trades? And why are they willing to lose so that you can win? Those two questions can go far in discerning which asset managers have skill and which are the patsies. Asset managers who don't have a succinct and well-reasoned answer to those questions probably do not have skill-based alpha.

Chasing Fads

The market has seen a growing body of specialized exchange-traded funds (ETFs) composed of stocks with specific traits designed to grab investors' attention and encourage fad chasing. Specifically, newly launched sector/industry or thematic ETFs tend to hold stocks that have "experienced recent price run-ups, had recent media exposure (especially positive exposure), had more positive earnings surprises, and displayed general traits that have been previously shown to indicate overvaluation" (Ben-David, Franzoni, Kim, and Moussawi 2021, p. 6). As a result, in sharp contrast to their prelaunch returns, these strategies on average deliver zero or even negative alpha.

Ben-David, Franzoni, Kim, and Moussawi (2021) examine essentially all ETFs that track model portfolios, whether indexes or not, and allocate them to four categories: broad index, smart beta, sector/industry, and thematic. The authors then ask the simple question: How did the model portfolios fare in the three years before the launch of the ETF and in the five years after, *measured relative to the benchmark selected by the asset manager*? This study looks at the model portfolio indexes, not the performance of the ETFs themselves. ETF performance will match the respective model portfolio, less fees and trading costs. The smart beta strategies (which include ideas that are probably truly smart, as well as many that are not)[6] typically added 2% a year in the three years before the strategy went live and nothing thereafter. Sector, industry, and thematic strategies typically added 3% to 5% a year before launch, then lost 4% to 5% a year thereafter—*before fees and trading costs*!

These results are consistent with the findings of Brightman, Li, and Liu (2015) who studied long-only index-tracking ETFs launched in the US market with at least a three-year record. Measuring the performance of the underlying indexes as reflected in Exhibit 1, they found that the average ETF delivered an excess return of 5% a year in the three-year period before launch and nearly zero in the three-year post-launch period. The excess return differences before and after the ETF launch are economically and statistically significant, validating their conclusion that ETF issuers launch products that largely track past winners.

Biotech, Y2K, B2B, social media IPOs, and VoIP are all examples of past fads that attracted vast sums of capital at high valuations. More recently, marijuana, psychedelics, crypto and NFT strategies, and SPACs, which collect assets for unspecified future acquisitions, have attracted significant capital. Apart from some of the most recent of those, most proved to be bubbles, which popped, destroying value for fad-chasing investors. The willing losers are the performance chasers, preferring the comfort of conformity over discomfort and profit.

This same issue afflicts the factor community in an interesting way. We—the factor investing community—typically combine factors that worked well in historical simulations and demonstrate that the combination has worked well in historical simulations. We then favor the factors that historically worked best, typically without regard for whether past alpha was largely a consequence of rising relative valuations or for whether current assets allocated to the strategies are sufficient to arbitrage away the structural alpha, net of trading costs and fees.

Nowcasting

Against a backdrop of constant noise and stimulation, the financial industry is also rife with nowcasting, which further engenders performance chasing. Media pundits, market prognosticators, and even investment boards frequently engage in nowcasting, which we refer to as the practice of explaining what has already happened *as if it is a forecast of the future*. Arnott and Bernstein (2002, 64) observed that “the investment management industry thrives on the expedient of forecasting the future by extrapolating the past.” Not only do such “predictions” rarely offer insight, they also invite us to chase trends, exacerbating that industry-wide problem.

Arnott and Treussard (2020) described nowcasting, pointing out why it is so very popular. Suppose we forecast what’s already happened by describing why it happened and use this description as a forecast for the future. What will people remember a year later? The forecast was “correct” up to the point at which it was offered. It was insightful in offering explanations for why something happened. If the previous trend persists in subsequent events, it will be recalled as prescient. If markets and circumstances reverse, it will be remembered as insightful and correct (up to the date it was offered). And, in the memory of the public, it will typically be incorrectly recalled as *preceding* the events that were “forecast,” even when the forecast comes after those events!

The capital markets discount current expectations, hence already reflecting the consensus past-is-prologue mindset. To be useful, a forecast should suggest how the future will differ from expectations and from the past. A nowcast does neither. Real forecasting is

[6] One of us (Arnott) has been called the “godfather of smart beta” in a number of publications. We think the “smart beta” label is fun but has been appropriated by the industry at large and attached to a host of ideas, many of them not at all smart. Accordingly, we think the label has become meaningless, except as a marketing tagline. Our reflections on the topic can be found on our website: https://www.researchaffiliates.com/home.

more difficult and much more dangerous to the prognosticator. A forecast that differs from the past will trigger a very different public recollection. If correct, it may be remembered as insightful and correct (or as reckless but lucky).[7] If incorrect, the public will correctly recall *when* the forecast was made because it differed from the antecedent market or economic conditions and will vividly remember the error.

That is why so few pundits forecast a future that differs markedly from the recent past. We find it useful, whether reading a newspaper or an academic article, to ask "Is this a forecast or a nowcast?" If the former, and if it is correct, it may have material value to an investor because current market prices may not reflect the insight. If it is a nowcast, it will be of no use if it is correct—the nowcast is already reflected in current prices—and will hurt us if it is not correct. We encourage our readers to try this exercise the next time you read an article that purports to offer insights into the future. We expect that you will be astonished at how many articles can be dismissed as nowcasting and can be safely ignored.

HOW CAN WE FIX THESE PROBLEMS?

We and many others have addressed these industry-wide practices repeatedly during the entire 30-year life of this journal. And yet, if anything, these practices are more prevalent now than ever before. Together, the practices condition us to dangerously overlook a few critical errors that are not entirely difficult to address. These include ignoring (1) the impact of revaluation alpha, (2) the effect of implementation costs and other hidden costs, and (3) our vulnerability to a performance expectations shortfall, each of which we explore next.

Revaluation Alpha

Instead of chasing the performance of recent winners, quants make the near-identical mistake of chasing the performance of recently successful quantitative models, factors, and strategies. Consider a strategy that has been gaining popularity and has gone from relative valuation levels similar to the broad market to a 25% relative valuation premium.[8] We call this change in valuation levels *revaluation alpha*. The consequence is that the performance of the strategy during the past five years would have been boosted by about 25%, leading to an illusion that the "alpha" of the strategy is huge. What does that tell us about the prospects for the strategy? At best, nothing. At worst, if mean reversion occurs in the relative valuation levels for the strategy, then lofty past performance may presage future underperformance. Investors ignore the impact of revaluation alpha at their peril.

[7] We are particularly fond of Keynes' observation on pages 157–158 in The *General Theory of Employment, Interest, and Money* (1936), typically—and unfortunately—only provided in its closing sentence:

> Finally, it is the long-term investor, he who most promotes the public interest, who will in practice come in for most criticism, wherever investment funds are managed by committees or boards or banks. For it is in the essence of his behaviour that he should be eccentric, unconventional and rash in the eyes of average opinion. If he is successful, that will only confirm the general belief in his rashness; and if in the short run he is unsuccessful, which is very likely, he will not receive much mercy. Worldly wisdom teaches that it is better for reputation to fail conventionally than to succeed unconventionally.

[8] We are deliberately using generic "relative valuation" terminology. While academia focuses on relative book/price, one might just as sensibly rely on relative sales to price, earnings to price, dividends to price, cash flow to price, or a blend of several metrics (our preferred approach). The point is that strategies, factors, anomalies, and quant models can come into or out of fashion, and in so doing, become more expensive or less expensive relative to the market.

In recent years, the gross profitability and low beta factors are examples of that phenomenon, and the value factor has been the counterexample.[9]

As Exhibit 2A shows in the gold line on the left scale, an investor in high-profitability businesses across the world's developed economies has, during the past 30 years, enjoyed 3.7 times the wealth accumulation relative to investors in low-margin businesses. That's terrific. But relative valuation multiples (using multiple metrics) for those stocks have more than doubled, from a relative valuation multiple of about 1.34 (meaning that high-margin businesses commanded a 34% premium over low-margin businesses in 1989) to a current relative valuation level of 2.69. Furthermore, the two lines in the exhibit are joined at the hip; their monthly movements have a beta of 0.89 and a correlation nearly as high at 0.82.[10]

If the relative valuation of high-margin stocks versus low-margin stocks was the same at yearend 2021 as it was in 1989, we can reasonably surmise that the cumulative alpha would have been cut in half to a modest 2% a year. Although 2% value-add sounds pretty good, it was produced by a leveraged 100% long/100% short strategy; a long-only strategy should capture roughly half of this value-add, or 1%. Furthermore, the value-add is measured before fees and trading costs. Investors' performance expectations, and perhaps even those of product vendors and those mentioned in the marketing literature, are often inflated by this revaluation alpha, which is presumably nonrecurring. Worse, future prospects may be compromised if relative valuations revert to historical norms.

The gap between the two lines is also informative. The "wedge" between them represents the alpha that is *not due to revaluation.* We call this *structural alpha*. This gap is remarkably stable—meaning there is approximately zero structural alpha *except for the changes in relative valuation*—for about three-fourths of the history, in the periods 1992–2000 and 2005–2021. One reasonable interpretation is that high-margin companies were outperforming only low-margin companies, *net of any revaluation alpha*, in the periods 1990–1991 and 2000–2004. This is not to say that profitability is a bad factor, only that half of its alpha came from revaluation. That said, we find it shocking that academe and the practitioner community pay scant attention to revaluation alpha. Indeed, many in both communities dismiss the idea that revaluation alpha matters.

Exhibit 2B shows similar results for the low beta factor. During the past 30 years, low beta stocks—using the Frazzini and Pedersen (2014) definition—have not outperformed high beta stocks in large-cap developed markets, although one could reasonably argue that in 30 years of mostly bull markets even a modest shortfall will look pretty good on a beta-adjusted basis. The fit is not nearly as good as the fit in Exhibit 2A for the gross profitability factor; the purple and gray lines have a beta of almost exactly 1.00, albeit with considerable "noise" in the relationship. Clearly, however, when low beta stocks are trading at a deep discount to high beta stocks, they do subsequently outperform, and when priced at a large premium, they do not.

[9] For purposes of this article, we are looking at large-cap stocks in the developed world index. The gross profitability factor is long the 30% of large-cap developed market stocks with the highest gross profitability (defined as revenue minus cost of goods sold/assets) and short the 30% of stocks with the lowest gross profitability. The low beta factor is long the 30% of large-cap developed market stocks with the lowest market beta (Frazzini and Pedersen 2014) and short the 30% of stocks with the highest market beta. The composite value factor is based on a price-to-fundamentals ratio that blends four measures of relative valuation: price-to-book value, price-to-five-year average sales, price-to-five-year average earnings, and price-to-five-year average dividends. The composite value factor is long the 30% of large-cap developed market stocks with the lowest price-to-fundamental ratios and short the 30% of stocks with the highest price-to-fundamental ratios.

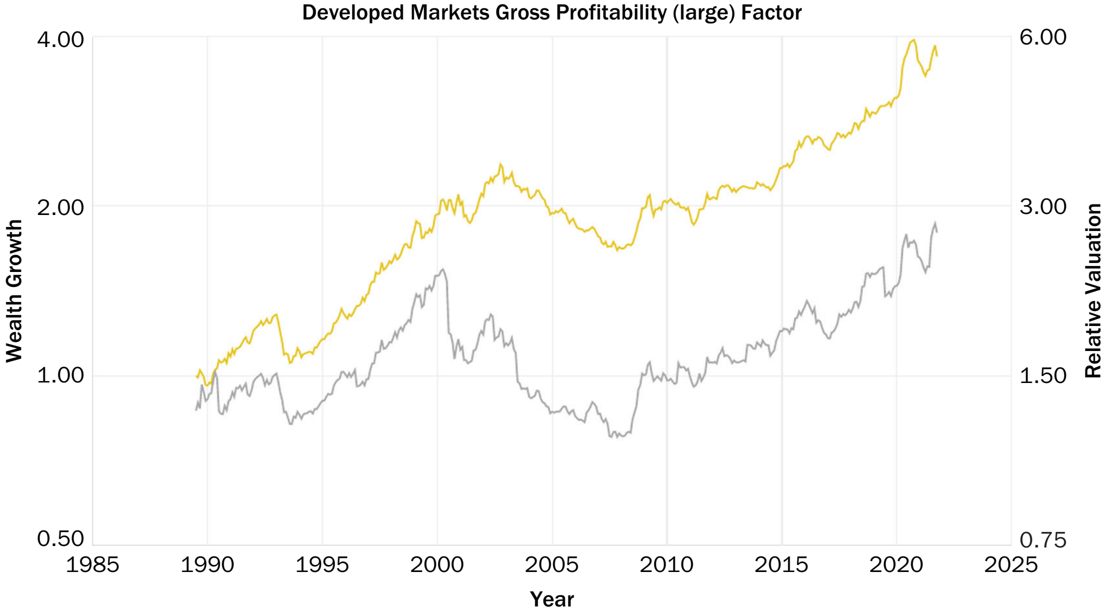
[10] The standard error for each of these betas is roughly 0.06, so we cannot reject a null hypothesis that the betas are 1.00. Indeed, this holds true for almost all factors on our Smart Beta Interactive website https://interactive.researchaffiliates.com/smart-beta#!/strategies. Even momentum, averaged across US, Developed, and Emerging Markets, Large and Small, has an average beta of 0.84.

Exhibit 2C shows the results for the composite value factor. As with gross profitability, the composite value factor shows a powerful link between relative valuation—by construction this is always at a discount relative to growth—and the performance of the factor. The beta is 1.09; valuation changes are slightly amplified in the relative performance. The exhibit illustrates that the underperformance of recent years is entirely due to a downward revaluation, to unprecedented cheapness for value relative to growth, which briefly eclipsed even the deep discounts at the peak of the tech bubble in

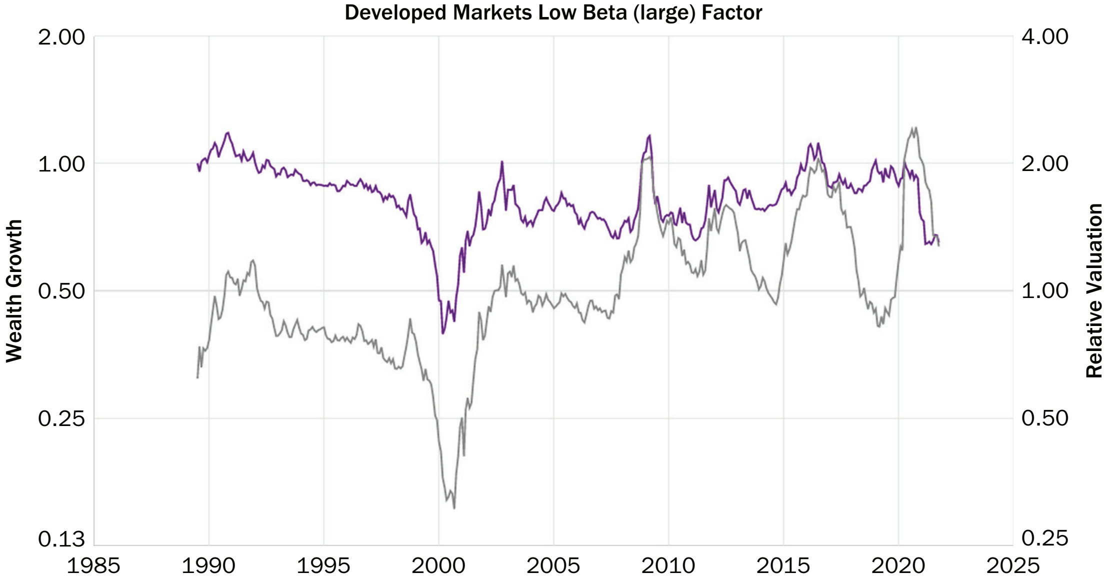
EXHIBIT 2

Illustrative Performance and Relative Valuation for Selected Factors, 1989–2021

Panel A: Gross Profitability Factor, Large-Cap Developed Market Stocks



Panel B: Low Beta Factor, Large-Cap Developed Market Stocks

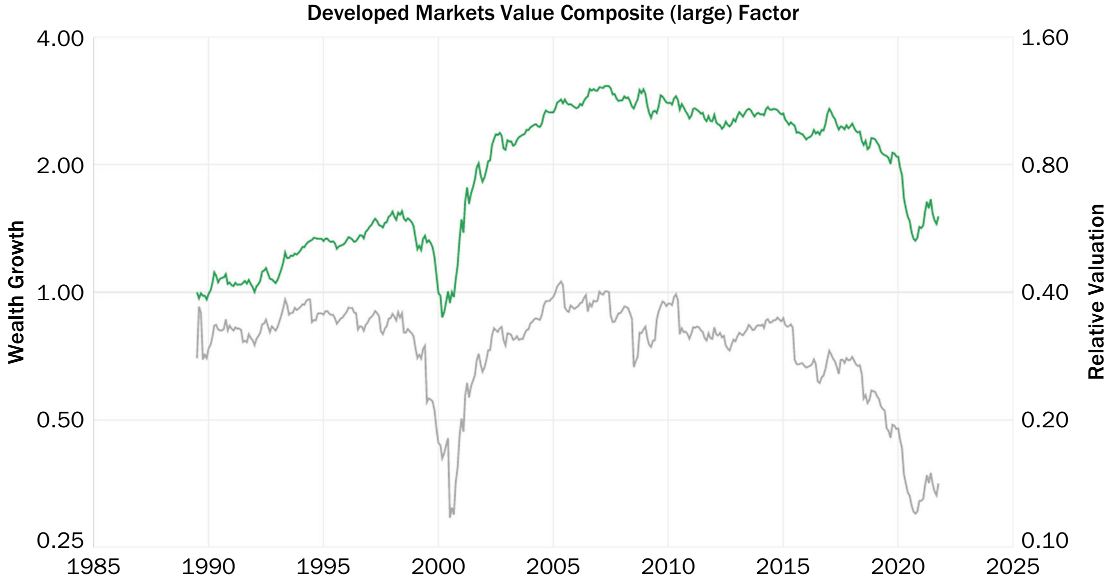


(continued)

EXHIBIT 2 *(continued)*

Illustrative Performance and Relative Valuation for Selected Factors, 1989–2021

Panel C: Composite Value Factor, Large-Cap Developed Market Stocks



NOTE: In Panel C, relative valuation for the composite value factor is a composite measure of four fundamental metrics: price-to-book value, price-to-five-year average sales, price-to-five-year average earnings, and price-to-five-year average dividends.

SOURCE: Research Affiliates, LLC, based on data from Worldscope and Datastream.

2000. The wedge between the green and gray lines is surprisingly reliable, widening whether value is winning or losing—and performance after deeply discounted valuations is often explosive. The wedge suggests a structural alpha of well above 4% a year, with value's bleak recent performance entirely explained—with room to spare—by a negative revaluation alpha, culminating in extraordinarily cheap valuations relative to growth. The underlying fundamentals for value stocks continue to improve relative to growth stocks, largely due to the migration effect documented by Fama and French (2007).

The article "How Can 'Smart Beta' Go Horribly Wrong?" (Arnott, Beck, Kalesnik, and West 2016) was greeted with aggressive condemnations by competitors, who seemed angered at the authors' suggestion that there were serious vulnerabilities in common applications of quantitative methods. We wonder if the competitors might have fared better in subsequent years if they had considered what might be right in the authors' message, rather than seeking angles of attack. Notably, in that article, the authors urged academia to demand that journal articles on new factors examine revaluation alpha so that academics are not feted for finding a "new factor" that merely "worked" by becoming more expensive. Had the article suggested that a stock that has doubled in price *and concurrently in relative valuation multiples* may not have had any structural alpha, no one would likely have found the article objectionable. We find it astonishing that revaluation alpha has not been subsequently explored in any serious way.[11]

[11] We would go as far as to suggest that the aversion to measuring revaluation alpha may be deliberate. Suppose an assistant professor finds a new factor through assiduous data mining with a *t*-stat of 4 and can show, through a Fama–French disaggregation of returns, for example, that the factor is materially different from the most popular factors. This is his or her shot at tenure, and perhaps even fame. Because no one publishes factors or strategies with negative alpha, the odds are pretty high that the factor has enjoyed positive revaluation alpha. If it is not de rigueur to measure the revaluation alpha, why should our assistant professor choose to run that test? The same logic applies to asset

Arnott, Beck, and Kalesnik (2016) carried out a thought experiment. Suppose one could go back in time to 1977 and offer three asset managers the formulas for eight of the more popular factors in use today, along with proof that these factors would all be profitable during the next 39 years (1977–2015, inclusive). One of the managers decides to equally weight the eight factors. The second decides to be clever, equally weighting the three best, choosing the factors with the best blended performance during the previous 1, 3, 5, and 10 years (our US factor return time series goes back to 1967). The third manager reasons that the best-performing factors in recent years are probably overvalued. Knowing that the factors all work on average during the next decades, the manager decides to equally weight the three *worst-performing* factors. It is safe to say that Manager 2 will garner the most assets, while Manager 3 will be dismissed as a reckless idiot. The naïve equal-weight manager earns a 39-year alpha of 2.4%. The manager who favors the "best" factors earns half as much, 1.2% a year, and the contrarian manager, favoring the "worst" factors, earns 3.3% a year.

Performance chasing afflicts even the most-savvy quants and hurts their clients.

Implementation Shortfall

Implementation shortfall refers to the difference between a paper portfolio's performance and the realized performance of a live portfolio. Many thematic funds and factor-based investment strategies are marketed according to backtested results estimated from paper portfolios. Many backtests do not respect portfolio management niceties, such as liquidity or trading costs. In a controversial 2020 article by Hou, Xue, and Zhang, 452 anomalies and factors were tested, with and without microcap stocks, the least liquid and smallest 2% of the stock market.[12] Of the 452 factors and anomalies, 65% (including 96% of the trading frictions category) were unable to clear the simple test hurdle of the absolute *t*-value of 1.96. Worse, 82% failed the higher multiple test hurdle of 2.78. Without the microcap stocks, the alphas and statistical significance were much smaller than originally reported, *for all of the 452 anomalies*.

It gets worse. Trading illiquid microcap stocks also incurs higher trading costs. The paper portfolios that have earned many academics their tenure ignore trading costs. By including thinly traded, illiquid stocks to bolster the backtest and then ignoring the likely trading costs that should result, the backtests can (and often do) exaggerate the alpha that investors might reasonably expect from a strategy. When used in marketing, this can inflate investor expectations, setting the stage for subsequent investor disappointment as most of the hoped-for alpha disappears into the pockets of brokers and market makers.

Various elements affecting a strategy's implementation cost include, but are not limited to, portfolio concentration, universe coverage, turnover, and capacity. Because some of these costs are hidden, not directly observable, or difficult to measure, they are often

managers who are seeking to launch and sell interesting new strategies. A new strategy, or a new tweak to a multi-factor strategy, may look brilliant on paper and therefore be highly marketable to investors. Again, selection bias means that the revaluation alpha was probably positive and perhaps explains a large chunk of past returns. Why would our investment manager wish to undermine future sales and profits by testing for a possible flaw in the strategy?

[12] The findings of Hou, Xue, and Zhang (2020) have been predictably attacked by the factor community, including by luminaries such as Richard Thaler. First, the article is mistitled as a replication; the authors do not really replicate because they use a common universe (with and without microcaps), thereby changing the method. Second, the authors have faced criticism because they expunge the very stocks in their research that are likely to drive the anomaly (less-liquid stocks). The critics overlook the fact that these factors and strategies cannot be costlessly replicated on an institutional scale. Trading costs in microcap stocks can easily devour the alpha. Even the classic Fama–French HML factor, which puts half its weight into small-cap names, which earn the lion's share of the HML alpha, suffers from this same problem.

ignored. These seemingly innocuous elements of product design, however, can lead to substantial differences in expected transaction costs. Disconcertingly, as the total assets managed by smart beta and factor strategies grow, implementation costs are contributing to a larger gap between backtested and live results. When we neglect small but critical details and costs, we set the stage for unhappy customers.

The Expectations Gap

While the return assumptions built into pension and retirement plans have steadily fallen in recent years, they may still be overoptimistic. According to the 2021 Milliman Corporate Pension Funding Study (Wadia, Perry, and Cook 2021), the average default long-term expected return of the 100 largest defined-benefit pension plans is 6.2%. The average investment return assumption for 77 public funds in fiscal year 2020, as surveyed by National Association of State Retirement Administrators (Brainard and Brown 2021), is 7.0%. Arnott (2020) showed a US public fund average of 7.2% as of yearend 2019. Pension plans' assumed returns are falling, albeit very slowly. We have long thought that such assumed metrics are subjective and incomplete, influenced by institutional pressures that encourage a high return assumption because the higher the assumed return, the better the funding ratio will seem and the lower the contribution that is needed.

Using an intuitive building-block framework based on the "building blocks"[13] of long-term return, we estimate the long-term nominal return of a pension fund to be 3.4%, or half of the projected rate that most public pensions expect. Our approach implies that the likelihood a pension fund can deliver 7% or more in the coming decade is under 1%. Those dismal prospects are not surprising when conventional bonds offer negative real yields and stocks are at near-record valuation multiples.

An actuary who signs off on high return expectations provides no guarantee the portfolio will deliver that high return. If our return assumption has any bearing on our future earned return, why not assume 20% and happily go home? The aspirational aspect is perhaps the biggest problem with the return assumption. Too many people believe it is reasonable to expect a target return merely because an accountant or actuary sets it as the return assumption. Scant attention is given to more objectively determined metrics, such as the risk-free funding ratio and official funding ratio, as a means to better assess a pension's state of health.

We believe that pension sponsors, whether public or corporate, should demand a more complete picture. How much of our promised benefits will our current assets fund if we simply immunize the portfolio against funding risk, with a duration-matched portfolio of long bonds and inflation-linked bonds? For the average public pension, that answer in 2020 was just under 30%, which means that the remaining 70% is presumed to come from premium returns, consisting of risk premiums relative to a default-risk-free duration-matched bond portfolio, and skill-based alpha. How much return would we need to earn on the current portfolio to serve our pension needs, without boosting our rate of pension contributions? For the average public pension fund in 2020, the average answer was just under 10%. This return is, in our view, essentially impossible to achieve from current market yields and valuation multiples. If we are correct, then states, counties, and cities (and corporate pension sponsors) will need to boost their pension contributions or reduce benefit payouts.

[13] For any investment, the long-term return is simply the sum of (1) the current yield, (2) the long-term growth rate of that income stream, and (3) the price change, plus or minus, of any changes in valuation levels. The current yield can be observed and the long-term historical real growth in an income stream is a useful basis for estimating potential future growth (though past performance is never a guarantee of future results). As for the changes in valuation levels, we have observed that multiples, yields, and spreads have had a powerful tendency to eventually mean revert toward historical norms.

To fund our pensions and other future spending needs, we need to *fund* them. It is dangerous to assume that the capital markets can fund our future spending without future contributions.

LESSONS LEARNED

Having reviewed the ways that investors succumb to these blunders, one may ask the question, *How can we manage these challenges?* In this concluding section, we propose a succinct three-part recommendation for quants, investors, and fiduciaries, based on observations spanning over, in the case of Arnott, a 40-year career. While these suggestions are not new and hardly exhaustive, they can serve as guiding principles to improve investor outcomes, both by delivering improved outcomes and by encouraging more sensible expectations.

Quants and Academic Researchers

Quants and academic researchers play a crucial role in setting appropriate investor expectations. When conducting backtests for either investment strategies or academic empirical work, they should impose a higher hurdle for declaring a backtest result, and they should willingly disclose their "negative" findings so that investors are fully aware of what has been tested and what did not work.

Fiduciary Professionals

Fiduciary professionals can help manage client expectations by reframing client performance reviews and institutionalizing contrarian behavior to encourage a mindset of long-termism. They can also help investors anticipate the exogenous risks (demographics, taxes, asset/liability mismatch, inflation, and so on) and help position clients' portfolios to minimize the expectations shortfall. One simple piece of advice they can offer clients would be to complement mainstream investments with diversifying asset classes, both to protect against environments that can devastate conventional holdings and to exploit attractive relative valuations.

Investors

Investors should recognize that in a performance backtest a meaningful portion of past returns may be tied to revaluation changes. So when assessing the forward-looking prospects of a strategy after an impressive run, ask whether the strategy is trading at cheap levels relative to history or whether recent strong performance may reduce or even reverse its future return prospects. Valuations matter! Investors shopping for factor strategies could consider, as a first step, using a framework to identify which factors are robust and to create structures to discourage the temptation to succumb to performance chasing. Last, but not least, hidden costs matter! Investors should ask whether a strategy's anticipated performance can be captured in the real world of trading costs and frictions.

REFERENCES

Arnott, R. 2020. "The COVID-19 Crash and the Abandonment of the Pensioner." Research Affiliates Publications (June).

Arnott, R., N. Beck, and V. Kalesnik. 2016. "Timing 'Smart Beta' Strategies? Of Course! Buy Low, Sell High." Research Affiliates Publications (September).

Arnott, R., N. Beck, V. Kalesnik, and J. West. 2016. "How Can 'Smart Beta' Go Horribly Wrong?" Research Affiliates Publications (February).

Arnott, R., and P. Bernstein. 2002. "What Risk Premium Is 'Normal'"? *Financial Analysts Journal* 58 (2): 64–85.

Arnott, R., C. R. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *The Journal of Financial Data Science* 1 (1): 64–74.

Arnott, R., and J. Treussard. 2020. "Forecasts or Nowcasts? What's on the Horizon for the 2020s?" Research Affiliates Publications (January).

Badrinath, S. G., and S. Wahal. 2002. "Momentum Trading by Institutions." *The Journal of Finance* 57 (6): 2449–2478.

Ben-David, I., F. Franzoni, B. Kim, and R. Moussawi. 2021. "Competition for Attention in the ETF Space." Fisher College of Business Working Paper No. 2021-03-001, Charles A. Dice Center Working Paper No. 2021-01, and Swiss Finance Institute Research Paper No. 21-03.

Black, F. 1986. "Noise." *The Journal of Finance* 41 (3): 529–543.

Brainard, K., and A. Brown. 2021. *Public Fund Survey*. National Associate of State Retirement Administrators (November). https://www.nasra.org/publicfundsurvey.

Brightman, C., F. Li, and X. Liu. 2015. "Chasing Performance with ETFs." Research Affiliates Fundamentals (November).

De Bondt, W., and R. Thaler. 1985. "Does the Stock Market Overreact?" *The Journal of Finance* 40 (3): 793–805.

Fama, E., and K. French. 2007. "Migration." *Financial Analysts Journal* 63 (3): 48–58.

Frazzini, A., and L. H. Pedersen. 2014. "Betting against Beta." *Journal of Financial Economics* 111 (1): 1–25.

Grinblatt, M., S. Titman, and R. Wermers. 1995. "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior." *American Economic Review* 85 (5): 1088–1105.

Harvey, C. R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72 (4): 1399–1440.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. "… and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

Hou, K., C. Xue, and L. Zhang. 2020. "Replicating Anomalies." *The Review of Financial Studies* 33 (5): 2019–2133.

Kahneman, D., and A. Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–454.

Keynes, J. M. 1936. *The General Theory of Employment, Interest and Money.* London: Macmillan Cambridge University Press.

McLean, R. D., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Tversky, A., and D. Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–232.

Wadia, Z., A. H. Perry, and C. J. Cook. 2021. *2021 Corporate Pension Funding Study*. Milliman. https://us.milliman.com/en/insight/2021-corporate-pension-funding-study.

Wermers, R. 1999. "Mutual Fund Herding and the Impact on Stock Prices." *The Journal of Finance* 54 (2): 581–622.