Accurate Benchmarking is Gone but Not Forgotten:
The Imperative Need to Get Back to Basics

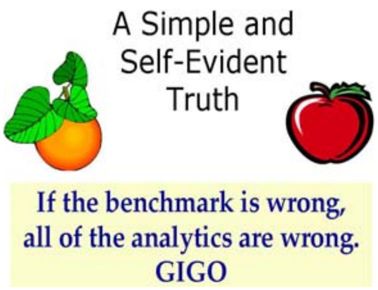
Synopsis

Investment performance evaluators have lost touch with a basic and self-evident truth: If the benchmark is wrong all of the analytics are wrong. The cost of this mistake is high because investment managers are hired and fired for the wrong reasons, sacrificing performance and fees. It's imperative that we get back to basics, that we get the benchmark right. Fiduciary prudence dictates best practice over common practice despite popular opinion to the contrary, as does the "do no harm" rule. Indexes and peer groups are the common forms of benchmarks. These are not best practices. The article describes how accurate benchmarks can be constructed from indexes and how peer group biases can be overcome.

Ronald J. Surz, 78 Marbella , San Clemente, CA 92673, 949/488-8339 (fax 0224), Ron@PPCA-Inc.com is president of PPCA, a software firm providing advanced analytics, primarily to financial consultants, for performance evaluation and attribution. Mr. Surz is also a principal of Risk-controlled Growth LLP, a fund-of-hedge-funds manager. He is active in the investment community, serving on numerous boards, including the CFA Institute's GIPS Investor/Consultant Subcommittee and After-tax Subcommittee, the Investment Management Consultants Association (IMCA) board of directors and standards committee, and several advisory boards of financial organizations. Mr. Surz earned an MS in Applied Mathematics from the University of Illinois and an MBA in Finance from the University of Chicago.

Accurate Benchmarking is Gone but Not Forgotten: The Imperative Need to Get Back to Basics

Ron Surz, PPCA Inc. and RCG LLC, 949/488-8339


A Simple and Self-Evident Truth
If the benchmark is wrong, all of the analytics are wrong. GIGO

Anything worth doing is worth doing well.
Anonymous

"Garbage in, garbage out" once was a primary concern of investment manager researchers. In recent years, however, this self-evident "GIGO" admonition has been largely forgotten. As the performance evaluation industry directed its focus to improving performance measurements, it lost sight of the basic and critical principle of accurate benchmarking. As a result, benchmarks are now routinely mis-specified, and performance measurements end up using a garbage input (i.e., an incorrect benchmark) to derive a garbage output, namely manager evaluations. The cost of this mistake is high; investment managers are hired and fired for the wrong reasons, thus sacrificing both performance and fees. It's imperative that we get back to basics, that we get the benchmark right. Fiduciary prudence dictates best practice over common practice despite popular opinion to the contrary.

Accurate benchmarking has recently become even more important due to the increasing interest in portable alpha. Alpha is transported by shorting the benchmark, thereby removing beta effects and leaving just the alpha. However, this can be done properly only if the benchmark is known—you can't short what you don't know. A related and equally important concern is that there is an alpha, which can be known only if the benchmark is properly specified. This is an insidious problem for hedge funds, where value added, or alpha, can easily be confused with factor exposures such as, for example, low market participation in a falling market.

To compound matters, a recent, oft-cited study finds that consultants are actually worse at picking managers than do-it-yourself investors. Bergstresser, Chalmers, and Tufano [2006], professors at Harvard Business School and the University of Oregon, documented that “financial intermediaries do a lousy job of allocating client assets to mutual funds.” Similarly, the press frequently observes that the average fund-of-hedge-funds consistently underperforms the average hedge fund, and that this underperformance is not due solely to fees. Simply stated, outside observers find that professionals have not delivered on their promise of finding skillful managers. The profession should heed this failure and take steps to change what has clearly been a losing game.

We now have an extensive menu of performance measurements (e.g., the Sharpe ratio, Sortino ratio, Treynor ratio, Information ratio, alpha), and newer, improved measurements—such as Omega Excess, Dr. Frank Sortino’s risk-and-style adjusted return, and the John Sturiale Consistency Ratio (SCR) which time and style adjust—continue to be introduced. Performance reports often include an array of such measures for the edification of sophisticated clients, but the truly sophisticated investor should be sure to substantiate the accuracy of the benchmark before trusting any measure of performance. If the benchmark is wrong all of the analytics are wrong. Today’s soup might be tasty, but the true gourmet cannot stomach it without knowing the ingredients.

That said, the current lack of focus on basics is understandable. Getting the benchmark right is complicated and more difficult than concocting new measurements or improving upon old ones. Tinkering with mathematical formulas is simply more fun than agonizing over the minutia of benchmark construction. Unfortunately, no amount of arithmetic can bail us out if the benchmark is wrong.

The two most common forms of benchmarks are indexes and peer groups. These are not best practices. In this article, we describe how accurate benchmarks can be constructed from indexes and how peer group biases can be overcome. Then we turn our attention to the unique challenges of benchmarking hedge funds. The article concludes with a discussion of accurate benchmarking for attribution analysis, which reveals the reasons for success and failure. Accurate benchmarking entails a lot of work, but it is well worth the effort. Just keep GIGO in mind—because we can forget about performance evaluation and attribution if we don't get the benchmark right.

Indexes

A benchmark establishes a goal for the investment manager. A reasonable goal is to earn a return that exceeds a low-cost, passive implementation of the manager's investment approach, because the investor always has the choice of active or passive management. The relatively recent introduction of style indexes helps, but these need to be employed wisely. Before style indexes were developed, there was wide acceptance and support for the concept of a "normal portfolio," which is a customized list of stocks with their neutral weights. "Normals" were intended to capture the essence of the people, process, and philosophy behind an investment product. However, only a couple of consulting firms were any good at constructing these custom benchmarks. Today we can approximate these "designer benchmarks" with style analysis, sometimes called "the poor man's normals." While style analysis may not be as comprehensive as the original idea of normal portfolios, it at least makes it possible for many firms to now partake in this custom blending of style indexes. Style analysis can be conducted with returns or holdings. Both approaches are designed to identify a style blend that—like normals—captures the people, process, and philosophy of the investment product.

Whether the returns or holdings approach to style analysis is used, the starting point is defining investment styles. The classification of stocks into styles leads to style indexes, which are akin to sector indexes such as technology or energy. It's important to recognize the distinction between indexes and benchmarks. Indexes are barometers of price changes in segments of the market. Benchmarks are passive alternatives to active management. Historically, common practice has been to use indexes as benchmarks, but style analyses have shown that most managers are best benchmarked as blends of styles. As a practical matter, we are no worse off with style blends, as the old practice is considered in the solution and there's always the possibility that the best "blend" is a single index.

One form of style analysis is returns-based style analysis (RBSA). RBSA regresses a manager's returns against a family of style indexes to determine the combination of indexes that best tracks the manager's performance. The interpretation of the "fit" is that the manager is employing this "effective" style mix because performance could be approximately replicated with this passive blend. Another approach, called holdings-based style analysis (HBSA), examines the stocks actually held in the investment portfolio and maps these into styles at points in time. Once a sufficient history of these holdings-based snapshots is developed, an estimate of the manager's average style profile can be developed and used as the custom benchmark. Note that HBSA, like normal portfolios, starts at the individual security level and that both normal portfolios and holdings-based style analysis examine the history of holdings. The departure occurs at the blending. Normal portfolios blend stocks to create a portfolio profile that is consistent with investment philosophy, whereas HBSA makes an inference from the pattern of point-in-time style profiles and translates the investment philosophy into style.

The choice between RBSA and HBSA is complicated and involves several considerations. Although RBSA has gained popularity, this doesn't necessarily mean that it's the best choice. The major trade-off between the two approaches is ease of use (RBSA) versus accuracy and ease of understanding (HBSA). RBSA has become a commodity that is quickly available and operated with a few points-and-clicks. Some websites offer free RBSA for a wide range of investment firms and products. Find the product, click on it, and out comes a style profile. Offsetting this ease of use is the potential for error. RBSA uses sophisticated regression analysis to do its job. As in any statistical process, data problems can go undetected and unrecognized, leading to faulty inferences. One such problem is multicollinearity, which exists when the style indexes used in the regression overlap in membership. Multicollinearity invalidates the regression and usually produces spurious results. The user of RBSA must trust the "black box," because the regression can't explain why that particular blend is the best solution. In his article that introduced RBSA, Nobel laureate Dr. William Sharpe [1988] set forth recommendations for the style indexes used in RBSA, known as the "style palette":

"It is desirable that the selected asset classes be:

- Mutually exclusive (no class should overlap with another)
- Exhaustive (all securities should fit in the set of asset classes)
- Investable (it should be possible to replicate the return of each class at relatively low cost)
- Macro-consistent (the performance of the entire set should be replicable with some combination of asset classes)."

The mutually exclusive criterion addresses the multicollinearity problem, and the other criteria provide solid regressors for the style match. The only indexes that currently meet all of these criteria are provided by Morningstar and Surz. Morningstar is available for U.S. stocks, while Surz indexes are provided for U.S., international, and global stock markets. Using indexes that do not meet Dr. Sharpe's criteria is like using low octane fuel in your high-performance car. See Picerno [2006] for an extensive discussion of a proper style palette.

Holdings-based style analysis (HBSA) provides an alternative to RBSA. The major benefits of HBSA are that the analyst can both observe the classification of every stock in the portfolio as well as question these classifications. This results in total transparency and understanding, but at a cost of additional operational complexity. HBSA requires more information than RBSA; that is, it needs individual security holdings at various points in time, rather than returns. Since these holdings are generally not available on the Internet, as returns are, the holdings must be fed into the analysis system through some means other than point-and-click. This additional work, sometimes called "throughput," may be too onerous for some, despite the benefits. Like RBSA, HBSA also requires that stocks be classified into style groups, or indexes. Dr. Sharpe's criteria work for both RBSA and HBSA; i.e., for consistency purposes, the same "palette" should be used for both types of style analysis. Note that the "mutually exclusive" and "exhaustive" criteria are particularly important in HBSA as it is highly desirable to have stocks in only one style group and to classify all stocks.

In certain circumstances, deciding between RBSA and HBSA is really a matter of Hobson's choice. When holdings data is difficult to obtain, as can be the case with some mutual funds and unregistered investment products such as hedge funds, or when derivatives are used in the portfolio, RBSA is simply the only choice. RBSA can also be used to calculate information ratios, which are style-adjusted return-to-risk measures. Some researchers are finding persistence in information ratios, so they should be used as a first cut for identifying skill. Similarly, when it is necessary to detect style drift or to fully understand the portfolio's actual holdings, HBSA is the only choice. Holdings are also required for performance attribution analysis that is focused on differentiating skill from luck and style--an important distinction. This level of analysis must use holdings because performance must be decomposed into stock selection and sector allocation. Returns cannot make this distinction.

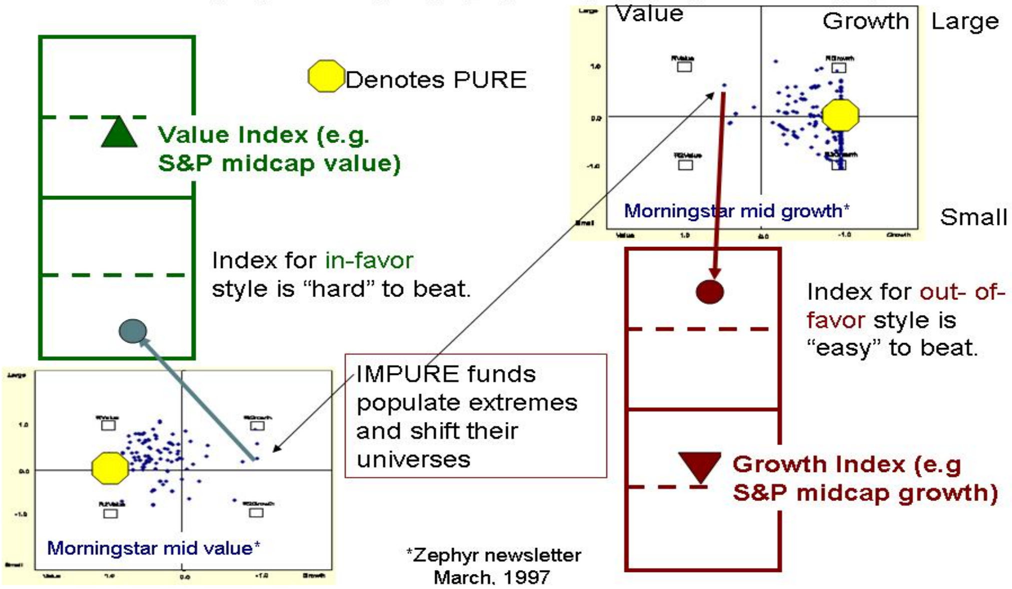
Custom benchmarks developed through either RBSA or HBSA solve the GIGO problem, but statisticians estimate that it takes decades to develop confidence in a manager's success at beating the benchmark, even one that is customized. This is because when custom benchmarks are used, the hypothesis test "Performance is good" is conducted across time. An alternative is to perform this test in the cross-section of other active managers, which is the role of peer group comparisons.

Peer Groups

Peer groups place performance into perspective by "ranking" it against similar portfolios. Accordingly, performance for even a short period of time can be adjudged significant if it ranks in the top of the distribution. When traditional peer groups are used, the hypothesis "Performance is good" is tested by comparing performance with that of a group of portfolios that is presumably managed in a manner similar to the portfolio that is being evaluated, so the hypothesis is tested relative to the stock picks of similar professionals. This makes sense, except that someone has to define "similar" and then collect data on the funds that fit this particular definition of similar. Each peer group provider has its own definitions and its own collection of funds, so each provider has a different sample for the same investment mandate. "Large cap growth" is one set of funds in one provider's peer group, and another set of funds in the next provider's peer group. These sampling idiosyncrasies are the source of the following well-documented peer group biases:

- *Classification* bias results from the practice of forcing every manager into a prespecified pigeonhole, such as growth or value. It is now commonly understood that most managers employ a blend of styles, so that pigeonhole classifications misrepresent the manager's actual style as well as those employed by peers. Classification bias is the reason that a style index ranks well, outperforming the majority of managers in an associated style peer group, when that style is in favor. Conversely, the majority of managers in an out-of-favor style tend to outperform an associated index. Until recently it was believed that skillful managers excelled when their style was out of favor. However, research has shown that this phenomenon is a direct result of the fact that many managers in a given style peer group are not "style pure," and it is this impurity, or classification bias, that leads to success or failure versus the index. See Hanachi [1998] and Surz [2006b] for more details. Exhibit 1 demonstrates the effect of classification bias. The scatter charts in this exhibit use RBSA to locate members of the Morningstar peer group in style space. As you can see, the tendency is for the funds to be somewhat similar, but significant compromises have been made.

Exhibit 1: Classification Bias



Denotes PURE

Value Index (e.g. S&P midcap value)

Index for in-favor style is "hard" to beat.

IMPURE funds populate extremes and shift their universes

Index for out- of-favor style is "easy" to beat.

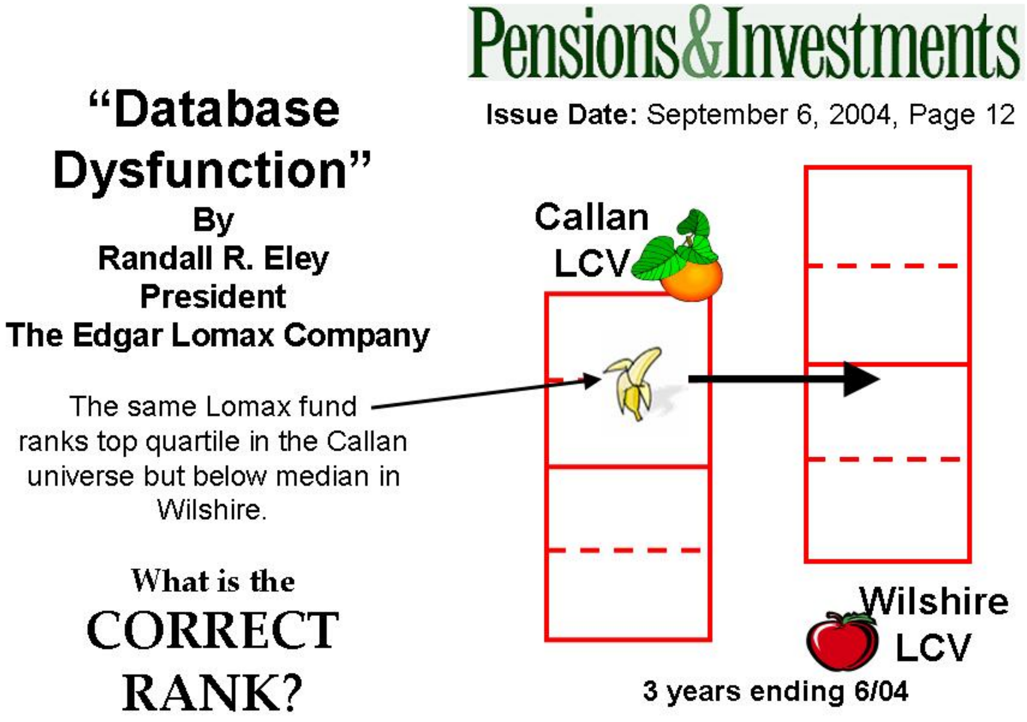
Growth Index (e.g S&P midcap growth)

*Zephyr newsletter March, 1997

Classification bias is a boon to client relations personnel because there is always an easy target to beat. When your style is out of favor, you beat the index; when it's in favor, you beat the median.

- *Composition* bias results from the fact that each peer group provider has its own collection of fund data. This bias is particularly pronounced when a provider's database contains concentrations of certain fund types, such as bank commingled funds, and when it contains only a few funds, creating a small sample size. For example, international managers and socially responsible managers cannot be properly evaluated using peer groups because there are no databases of adequate size. Composition bias is the reason that managers frequently rank well in one peer group, but simultaneously rank poorly against a similar group of another provider, as documented by Eley [2004]. Don't like your ranking? Pick another peer group provider. It is frequently the case that a manager's performance result is judged to be both a success and a failure because the performance ranks differently in different peer groups for the same mandate, such as large cap value. Exhibit 2 summarizes the Eley article.

Exhibit 2: Composition Bias



Pensions&Investments

Issue Date: September 6, 2004, Page 12

"Database Dysfunction"

By

Randall R. Eley

President

The Edgar Lomax Company

The same Lomax fund ranks top quartile in the Callan universe but below median in Wilshire.

What is the CORRECT RANK?

- *Survivorship* bias is the best understood and most documented problem with peer groups. Survivor bias causes performance results

> to be overstated because defunct accounts, some of which may have underperformed, are no longer in the database. For example, an unsuccessful management product that was terminated in the past is excluded from current peer groups. This removal of losers results in an overstatement of past performance. A related bias is called "backfill bias," which results from managers withholding their performance data for new funds from peer group databases until an incubator period produces good performance. Both survivor and backfill biases raise the bar. A simple illustration of the way survivor bias skews results is provided by the "marathon analogy," which asks: If only 100 runners in a 1,000-contestant marathon actually finish, is the 100th runner last? Or in the top 10%?

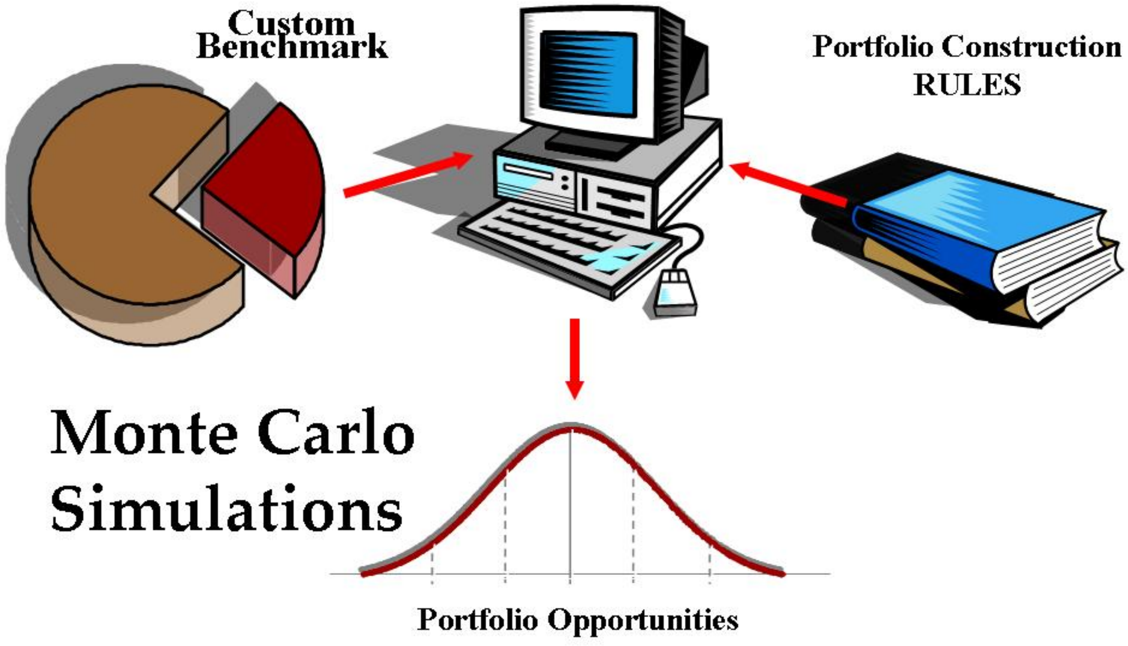
In summary, peer group comparisons are more likely to mislead than to inform, and therefore they should be avoided. Given the common use of peer groups, we realize this position is an unpopular one, but sometimes common practice defies common sense. (Think cigarettes.) These bias problems are not solved by finding the "right peer group." Try as we may, there is no way to make the biases described above go away. The most that can be done is to try to minimize the effects of these biases, which can best be accomplished with the approach described in the next section.

Unifying Custom Benchmarks with Peer Groups

Let's summarize what we've covered so far. Custom blended indexes provide accurate benchmarks, but we have to wait decades to gain confidence in a manager's success at beating the benchmark. Peer groups don't have this "waiting problem," but are contaminated by myriad biases that render them useless. A solution to these problems is actually quite simple, at least in concept, but was only recently made practical when the requisite computing power became available. The solution uses custom benchmarks to create a peer group backdrop that does not have a waiting problem, that is, we know right away if a manager has significantly succeeded or failed.

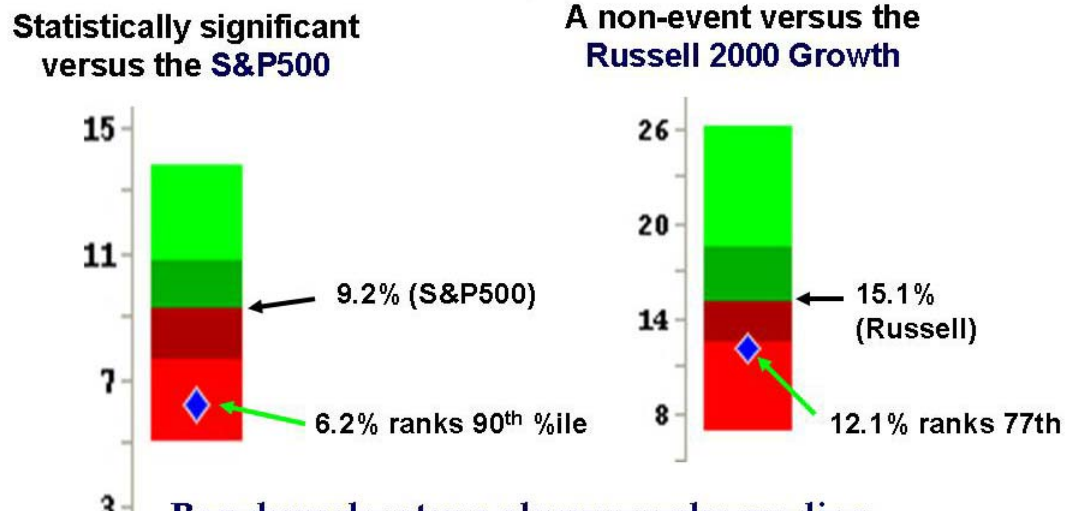
As noted above, performance evaluation can be viewed as a hypothesis test that assesses the validity of the hypothesis "Performance is good." To accept or reject this hypothesis, we construct an approximation of all of the possible outcomes and determine where the actual performance result falls. This solution begins with identification of the best benchmark possible, like a custom index blend, and then expands this benchmark into a peer group by creating thousands of portfolios that could have been formed from stocks in the benchmark, following reasonable portfolio construction rules. This approach, illustrated in Exhibit 3, combines the better characteristics of both peer groups and indexes, while reducing the deficiencies of each.

Exhibit 3: Unifying Benchmarks with Peer Groups



Importantly, statistical significance is determined much more quickly with this approach than with benchmarks because inferences are drawn in the cross-section rather than across time. In other words, the ranking of actual performance against all possible portfolios is a measure of statistical confidence. Exhibit 4 demonstrates this determination. Let’s say the manager has underperformed the benchmark by 3%. The exhibit shows that in a recent quarter, this underperformance would have been significant if the S&P 500 were the benchmark, but not significant if the benchmark were the Russell 2000 growth. We use 90% confidence as the breakpoint for declaring significance. Because they provide indications of significance very quickly, Monte Carlo simulations (MCS), as the approach is known, solve the waiting problem of benchmarks.

Exhibit 4: Significance of missing the benchmark by 3%



Benchmark return always ranks median.

In the due diligence process, there are two central questions: What does this manager do (style, etc.)? and Does the manager do this well? The first question addresses the form of the investment, and the second identifies the substance, or skill. In this context, the benchmark provides the answer to the first question: What does this manager do? The ranking within the manager’s customized opportunity set answers the second question “Does the manager do this well?”

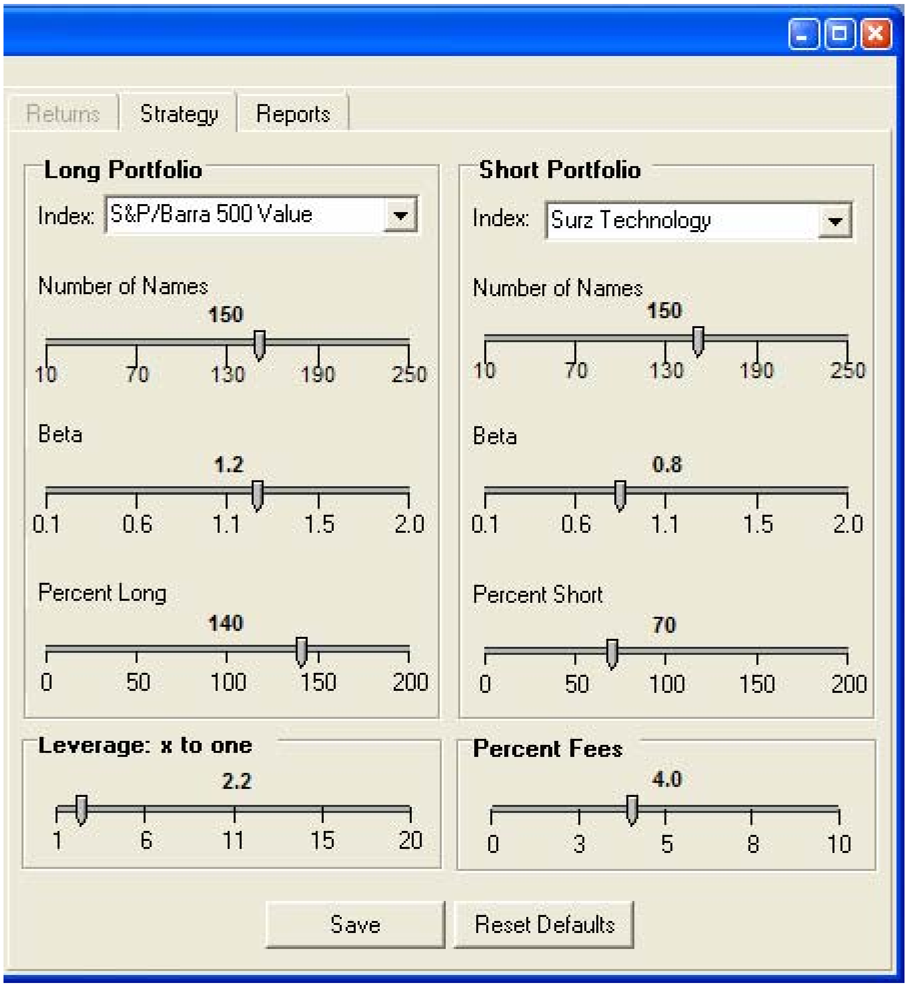
Note that in properly constructed Monte Carlo simulations, the benchmark always ranks median. See Sharpe's "The Arithmetic of Active Management" [1991] for an explanation why this must be the case. This provides for the interpretation of an MCS ranking as the "statistical distance" of return away from the benchmark.

The Monte Carlo simulation approach is not new. MCS has been used to evaluate traditional investing for more than a decade (see Surz [2006], Burns [2004], and Bridgeland [2001]). Even though MCS has not yet been accepted as standard practice (see Chernoff [2003] and Picerno[2003]), this doesn't mean that the idea is faulty. Modern Portfolio Theory (MPT), for example, took 30 years to become established. Further improving the potential for acceptance, MCS technology has been extended to hedge funds, where recognition of the fact that peer groups don't work for performance evaluation has lowered inherent barriers to comprehension and adoption, as described in the next section.

Hedge Funds

The first question of due diligence—"What does this manager do?"—is typically difficult to answer in the hedge fund world. However, a basic tenet should be kept in mind: Don't invest in something you don't understand. Work being conducted in returns-based analysis of hedge funds helps to answer this first question about the form of the investment. See, for example, Fung and Hsieh [2003], who demonstrate that the beta of a specific hedge fund can be replicated with a long-short blend of passive portfolios such as exchange-traded funds (ETFs). We shouldn't pay for beta, but its identification sets the stage for the second question regarding substance. As with traditional long-only investing, Monte Carlo simulations provide the answer to the question of manager skill. In constructing a specific custom peer group, Monte Carlo simulations follow the same rules that individual hedge fund managers follow in constructing portfolios, going both long and short, following custom benchmark specifications on both sides, as well as using leverage, employing controls such as those shown in Exhibit 5.

Exhibit 5: Sample Control for Hedge Fund MCS



Long Portfolio
Index: S&P/Barra 500 Value
Number of Names 150
Beta 1.2
Percent Long 140

Short Portfolio
Index: Surz Technology
Number of Names 150
Beta 0.8
Percent Short 70

Leverage: x to one 2.2

Percent Fees 4.0

MCS addresses the uniqueness challenge of evaluating hedge fund performance by randomly creating a broad representation of all of the possible portfolios that a manager could have conceivably held following his unique investment process, thereby applying the scientific principles of modern statistics to the problem of performance evaluation. This solves the major problem of hedge fund peer groups documented by Kat [2003], i.e., the members of hedge fund peer groups are uncorrelated with one another, violating the central homogeneity principle of peer groups. Some observers say it's good that the members of hedge fund peer groups are unlike one another, because this produces diversification benefits. While it may be good for portfolio construction, it's bad for performance evaluation. Funds in hedge fund peer groups should not be compared with one another because it's like comparing apples and oranges. Hedge funds really do require not only custom MCS peer groups for accurate evaluation, but also custom benchmarks that reveal both the longs and shorts, thereby estimating the hedge fund's beta. A ranking in a hedge fund MCS universe renders both the alpha and its significance.

Performance Attribution

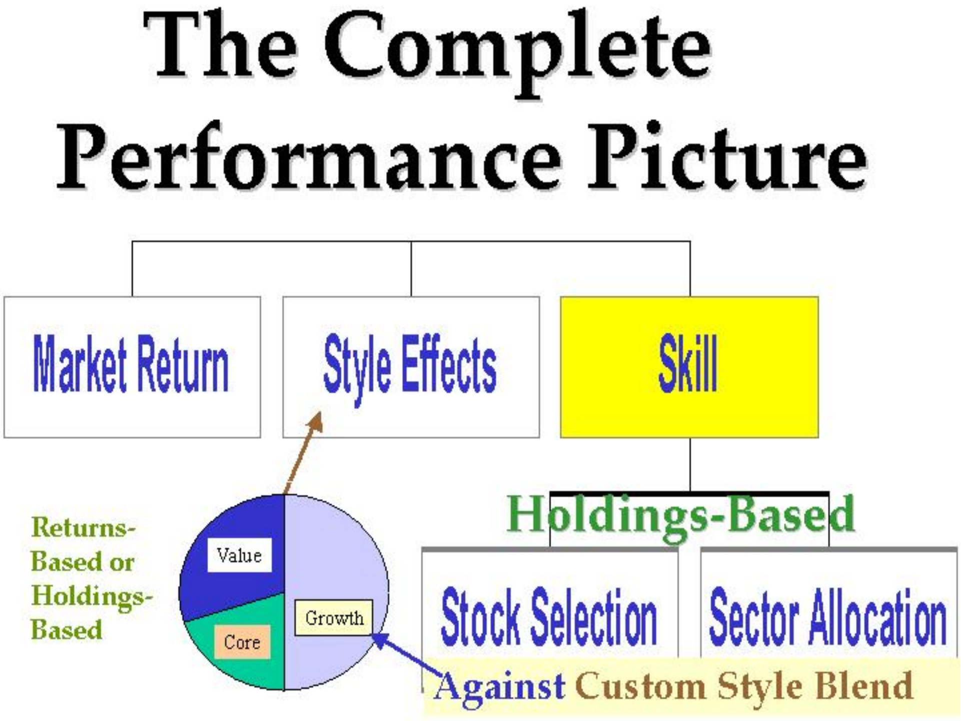
Up to this point we have been discussing performance evaluation, which determines whether performance is good or bad. The next, and more crucial, question is "Why?", which is the role of performance attribution. Attribution is important because it is forward-looking, providing the investor with information for deciding if good performance is repeatable in the future. We want to know which sectors had good stock selection and/or favorable allocations and if the associated analysts are likely to continue providing these good results. We also want to know what mistakes have been made and what is being done to avoid these mistakes in the future. These are important considerations that fortunately can be addressed with the same accurate, customized benchmark that we've described for use in performance evaluation.

This practice enables us to steer clear of the problem associated with more common attribution systems, i.e., the frequent disconnect between the benchmark used for evaluation and the one used for attribution. This disconnect is due to the fact that most performance attribution systems are currently limited to popular indexes and cannot accommodate custom benchmarks. This unfortunate limitation creates the very GIGO problem we've set out to avoid. We should not throw away all of our hard work in constructing an accurate benchmark when it comes to the important step of attribution. Put another way, we shouldn't bother with attribution analyses if we can't customize the benchmark. We'll just spend a lot of time and money to be misled and misinformed.

Conclusion

Getting back to basics is more than just a good thing to do. Getting the benchmark right is a fiduciary imperative, an obligation. Even if you don't agree with this article's recommended best practices, you can't deny the failure of common practices. Something has to change. Current common practices are not best practices; we can and must do better.

The components of investment return as we understand them today are summarized in the accompanying graphic entitled "The Complete Performance Picture." The new element in this picture, beyond Modern Portfolio Theory, is indicated by the box labeled "Style Effects." MPT, which relies exclusively on market-related effects, has not worked as predicted because of the powerful influences of investment style. It's easy to confuse style with skill, but difficult to make good decisions once this mistake has been made. Accurate benchmarks are customized to each individual manager's style and should be used for both performance evaluation and performance attribution. Monte Carlo simulations expand these custom benchmarks into accurate and fair universes, similar to peer groups but without the biases, and provide indications of significance very quickly. Both traditional and hedge fund managers are best reviewed with these techniques.



ENDNOTE

Many thanks to Gale Morgan Adams for her remarkable editing.

REFERENCES

Bergstresser, Daniel B., Chalmers, John M.R., and Tufano, Peter. "Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry." American Finance Association (AFA), 2006 Boston Meetings (January 16, 2006). Forthcoming available at SSRN: http://ssrn.com/abstract=616981

Bridgeland, Sally. "Process Attribution – A New Way to Measure Skill in Performance Construction." *Journal of Asset Management*, December 2001.

Burns, Patrick. "Performance Measurement via Random Portfolios." Newsletter for Professional Risk Managers International Association (PRMIA), December 2004.

Chernoff, Joel. "Consultant Offers a Way to End Classification Bias." *Pensions & Investments*, August 18, 2003, page 3.

Eley, Randall R. "Database Dysfunction." *Pensions & Investments*, September 6, 2004, page 12.

Hanachi, Shervin. "Can the Average U.S. Equity Fund Beat the Benchmarks?" *Journal of Investing*, Summer 2000.

Kat, Harry M. "10 Things That Investors Should Know About Hedge Funds." *The Journal of Wealth Management,* Vol. 5, No. 4, Spring 2003, pp 72-81.

Picerno, James. "In the Quest to Identify Investment Skill, Ron Surz Claims He Has the Better Mousetrap." *Bloomberg Wealth Manager*, June 2003, pp 80-82.
_____________. "A Style All His Own." *Wealth Manager*, September 2006, pp 64-66.

Sharpe, William F. "Determining a Fund's Effective Asset Mix." *Investment Management Review*, December 1988, pp 59-69.
____________. "The Arithmetic of Active Management." *The Financial Analysts Journal,* Vol. 47, No. 1, January/February 1991, pp 7-9.

Surz, Ronald J. "A Fresh Look at Investment Performance Evaluation: Unifying Best Practices to Improve Timeliness and Accuracy." *The Journal of Portfolio Management,* Summer 2006, pp 54-65.
___________. "A Common Misperception about Beating The Index." *White Paper,* August 2006.