THE LIMITS OF HISTORY

Anyone who's ever read a prospectus knows that "past performance is no guarantee of future results." The sentence has almost assumed the status of a cliché, but one of the less-appreciated properties of clichés is that they're true.[1]

When a prospectus says "past performance," it means the actual, real-time historical performance of the security or investment strategy in question. But history is not the only source of past performance data. Performance data can also be simulated, or backtested, and simulated data raise issues of their own. A recent paper[2] addressed these issues in the context of what are typically called "strategy" or "factor" indices. Such indices are designed not simply to replicate the performance of an asset class, but rather to reflect the performance of a factor or attribute with which a particular return pattern is thought to be associated.

Strategy indices are typically introduced to the market by showing backtested data—by producing, in other words, a hypothetical or simulated record of how the strategy would have performed historically. (This is possible only if the index is rules-based, quantitative and contains no look-ahead bias.) The implicit assumption is that the backtest will be a fair indicator of the strategy's real-time performance. It is precisely this assumption that Dickson et al. question, demonstrating that backtests predict the presence of excess returns far more frequently than real-time investments actually produce excess returns.

The conclusion that simulations should be taken with a grain, and occasionally a great many grains, of salt is surely unexceptionable. Anyone being asked to invest capital on the basis of a backtest will be well served by a healthy dose of skepticism. And yet, when a new idea arises, simulated performance may be the only performance there is. "Not every attractive investment opportunity can be sold complete with the security blanket of a five-year performance history.… While it is occupationally healthy to be skeptical of products that exist only on paper and promises, such skepticism should not extend beyond the point where sound opportunities are missed."[3] A potential investor who refuses to look at a backtest is effectively refusing to consider most new ideas.

The point is not, we submit, that an investor should never trust a backtest. Rather, the investor's goal should be to understand what it is about backtests that makes them more or less trustworthy. At its most fundamental level, **this is primarily a question not about backtests, but about historical data generally, both simulated and real**—a question about the limits of historical

Contributor:

Craig J. Lazzara, CFA
Senior Director
Index Investment Strategy
craig_lazzara@spdji.com

[1] See, e.g., Aye Soe and Frank Luo, "Does Past Performance Matter? The Persistence Scorecard," www.spindices.com/documents/spiva/persistence-scorecard-december-2012.pdf .
[2] Joel M. Dickson, Ph.D. et al., "Joined at the hip: ETF and index development," July 2012. See https://advisors.vanguard.com/iwe/pdf/ICRJAH.pdf?cbdForceDomain=true.
[3] John D. Freeman, "Behind the Smoke and Mirrors: Gauging the Integrity of Investment Simulations," *Financial Analysts Journal*, November-December 1992, pp 26-31.

data, regardless of their source. We suggest three factors that investors should consider in their evaluation of historical data on new index concepts:

- Plausibility in index design
- Clarity of index objective
- Sensitivity to market regime

PLAUSIBILITY IN INDEX DESIGN

Strategy indices are typically driven by implicit or explicit economic models—by the notion, in other words, that X factor or factors (the explanatory variables) will often be associated with Y outcome. We speak of a model as *plausible* if its explanatory variables have a reasonable connection to the outcome we're trying to explain. **A model that lacks plausibility should be rejected out of hand.** This is an obvious point, but given the size of available databases and the power of computers, it's entirely possible to identify spurious independent variables that seem to have significant "explanatory" power. As an example of this phenomenon, consider that the production of butter in Bangladesh does a remarkable job of "forecasting" the level of the S&P 500®.[4]

In the case of backtested data, this problem is compounded by what academics call information asymmetry—otherwise said, the creator of the simulation knows a lot more about it than the potential buyer does.[5] Backtest developers have large amounts of data at their disposal, as well as an impressive array of econometric tools. And, of course, they're human beings, which means that they share the same range of virtues and defects that we all share. Some are strikingly brilliant, some are of more ordinary intelligence. Some are extraordinarily careful and detail-oriented, some are more casual. Some are paragons of honesty and integrity, some are... not. And it's typically the case that if you torture the data enough, you can make them talk.[6] How can an investor have confidence that a set of promising backtest results represents genuine investment opportunity and not a coerced confession? A good place to start is to ask whether the relationships underlying the backtest are intuitively plausible.

As in many things, when thinking about plausibility, Occam's razor is a useful guide. The simpler and more transparent the backtest, the easier it is for a potential user to replicate it and to verify that the backtest is an accurate simulation of what might have been historically achievable. The more complex and opaque the backtest, the more difficult it is to verify. This is especially true if the backtest includes any proprietary elements—i.e., procedures or formulae that the backtester is not willing to disclose fully.

The arguments for simplicity and transparency amount, in effect, to saying that a simpler backtest makes it harder for the backtester to fool a potential client. Just as importantly, a simpler backtest makes it harder for the backtester to fool himself. A complicated, data-intensive model can be more easily overfit than a simpler formulation. "Overfit," in this context, means simply that the model includes variables that explain past behavior well but are useless in predicting future behavior. An overfit model can produce precisely the result we're trying to avoid—a backtest that looks good historically but falters when applied in real time. The simpler and more plausible the backtest, the less likely it is that the backtester will fall into this trap.

CLARITY OF INDEX OBJECTIVE

By *clarity*, we mean that a strategy index should have a clearly defined goal. Another way to express this point is to say that **the index developer should know, before he begins, how he'll know he's finished**. And the finish line *cannot* be defined simply as the point at which the backtest results outperform vanilla. This would be a perfectly good, indeed expected, answer from an active manager, but index providers must embrace a different standard.

A good way to think about building and testing a strategy index is to start by defining the *pattern of returns* that the index is intended to reflect.[7] For example, the S&P 500 Low Volatility Index was

[4] David J. Leinweber, "Stupid Data Miner Tricks: Overfitting the S&P 500," 1995. See http://shookrun.com/documents/stupidmining.pdf.
[5] In that sense, relying on a backtest is like buying a used car.
[6] See Freeman, *op. cit.* "The ruthless energy of computers means researchers can get blood from a turnip."
[7] Notice that the active manager looking for outperformance may not care much about the *pattern* of returns, although he will care a great deal about the *level* of returns.

designed to reflect a pattern of returns that is less volatile than the returns of the parent S&P 500.[8] In some market environments it might also produce higher returns than the parent index, but that's not an essential element of its design. Otherwise said, the index "succeeds" when it delivers a less volatile return pattern than its parent, not when it outperforms its parent.

In contrast, the S&P 500 High Beta Index[9] was designed to reflect higher levels of systematic risk than the parent S&P 500, i.e., to magnify moves in the parent index in both up and down markets. In some market environments this might produce a performance advantage, but again, that's not an essential element of its design—in this case, the index "succeeds" when it magnifies the parent's moves, regardless of direction.

Historical data—whether backtested or live—gain credibility when the index developer can identify his objective clearly, and when the pattern of historical returns matches that objective.

SENSITIVITY TO MARKET REGIME

The final criterion, and perhaps the most important one, requires the consumer of historical performance data to understand the nature of the market environment in which that performance was generated—and how that environment's influence might affect future results. This requires us to do some explicit analysis of how strategy indices react to the market environment in which they operate.

Consider the following example of two (actual) strategy indices, both of which are subsets of the S&P 500. The indices were developed using backtested data spanning 1991 through 2010.

Exhibit 1: Backtested Performance of Two Strategy Indices Versus the S&P 500 (1991-2010)

	Compound Annual Return 1991–2010 (%)	Annual Value Added (%)	Average Monthly Return (%)	Percentage of Months Outperforming the Benchmark (%)
Index A	10.17	1.03	0.87	49
Index B	9.48	0.34	1.13	50
S&P 500	9.14	NA	0.83	NA

Source: S&P Dow Jones Indices LLC. Data as of Dec. 31, 2010. Index performance based on total return USD. Charts and graphs are provided for illustrative purposes. Past performance is no guarantee of future results. These charts and graphs may reflect hypothetical historical performance. Please see the Performance Disclosure at the end of this document for more information regarding the inherent limitations associated with backtested performance. Value Added represents the difference between the performance of Index A or Index B and that of the benchmark S&P 500.

In the backtest period, both indices outperformed the vanilla S&P 500. Moreover, their outperformance didn't depend on a relatively short burst of spectacular returns; both indices outperformed in roughly half the months of the backtest period.

Both Index A and Index B went live in early 2011, so actual 2011-2012 data let us check how well the 20-year backtest corresponded with live performance. And here, the results are—at least on first glance—alarming. In 2011, Index A did very well but Index B seems to have fallen apart completely:

Exhibit 2: Actual Performance of Two Strategy Indices Versus the S&P 500 (2011)

	Total Return 2011 (%)	Value Added (%)	Average Monthly Return (%)	Percentage of Months Outperforming the Benchmark (%)
Index A	14.78	12.67	1.18	67
Index B	-18.01	-20.12	-1.24	25
S&P 500	2.11	NA	0.27	NA

Source: S&P Dow Jones Indices LLC. Data as of Dec. 31, 2011. Index performance based on total return USD. Charts and graphs are provided for illustrative purposes. Past performance is no guarantee of future results. Value Added represents the difference between the performance of Index A or Index B and that of the benchmark S&P 500.

[8] More information is available at www.spindices.com/indices/strategy/sp-500-low-volatility-index.
[9] See www.spindices.com/indices/strategy/sp-500-high-beta-index for more detail.

Index B's 2011 performance, in fact, looks like a textbook illustration of the phenomenon we're trying to avoid: an index that looks perfectly reasonable in backtesting but fails to deliver in real time.

The tables turned in 2012. Index B recovered from its embarrassing performance in 2011, while Index A fell short of our backtest-driven expectations:

Exhibit 3: Actual Performance of Two Strategy Indices Versus the S&P 500 (2012)

	Total Return 2012 (%)	Value Added (%)	Average Monthly Return (%)	Percentage of Months Outperforming the Benchmark (%)
Index A	10.30	-5.70	0.83	42
Index B	18.20	2.20	1.58	58
S&P 500	16.00	NA	1.29	NA

Source: S&P Dow Jones Indices LLC. Data as of Dec. 31, 2012. Index performance based on total return USD. Charts and graphs are provided for illustrative purposes. Past performance is no guarantee of future results. Value Added represents the difference between the performance of Index A or Index B and that of the benchmark S&P 500.

What are we to make of these results? Did Index A simply have a delayed reaction to an overfit backtest? Was Index B just the victim of bad luck in 2011? One response, of course, would be to argue that the live results of both indices confirm every suspicion we ever had about the reliability of backtests—clearly both Index A and Index B behaved much differently in real time than the backtest led us to believe. Or did they?

CONTINGENT PERFORMANCE ANALYSIS

What if the relative performance of Index A and Index B was influenced by factors that operated differently in the backtest period than in the live period? If both Index A and Index B are importantly driven by some aspect of the market regime in which they operate, and if the regime during the live period is different from the regime in the backtest period, then we have another prism through which to examine their performance.

There are many possible ways to identify an economic or market regime; one of the simplest (at least when dealing with U.S. equity markets) is to look at the returns of the S&P 500. In the table below, we partitioned our backtested data into "poor" months (when the S&P 500 returned -4.0% or worse), "good" months (when the S&P 500 returned +4.0% or better), and "middling" months (the other ones).[10]

Exhibit 4: Backtested Performance of Two Strategy Indices in Various Market Environments (1991-2010)

		Average Total Return (%)			Average Value Added (%)		Probability of Outperformance (%)	
Months	Number of Months	S&P 500	Index A	Index B	Index A	Index B	Index A	Index B
Poor	30	-7.2	-3.4	-13.2	3.8	-6.1	97	3
Middling	158	0.6	0.7	0.8	0.1	0.2	52	48
Good	52	6.2	3.9	10.4	-2.3	4.2	13	81
All	240	0.8	0.9	1.1	0.0	0.3	49	50

Source: S&P Dow Jones Indices LLC. Data as of Dec. 31, 2010. Index performance based on total return USD. Tables are provided for illustrative purposes. Past performance is no guarantee of future results. These charts and graphs may reflect hypothetical historical performance. Please see the Performance Disclosure at the end of this document for more information regarding the inherent limitations associated with backtested performance. Value Added represents the difference between the performance of Index A or Index B and that of the benchmark S&P 500.

In our 20-year backtest, there were 30 months when the S&P 500 fell by more than 4%. In those months, the average return of Index A was -3.4%, a 3.8% improvement over the S&P 500's -7.2% average return. Index B, on the other hand, underperformed dramatically in the poor months, lagging the S&P 500 by an average of 6.1%. If we examine not the difference in returns but the frequency with which Indices A and B outperformed the S&P 500, the results are consistent: Index A was highly likely to outperform the S&P 500 in the poor months, while Index B was highly likely to lag.

[10] The choice of ±4% bands is arbitrary, but tightening or loosening them makes very little difference—the numbers in the chart would change, but the conclusion would not.

We see more or less the opposite results in the months of "good" performance (i.e., when the S&P 500 was up by 4% or more). In those months, Index A lagged consistently and Index B was highly likely to outperform (by an average of 4.2% monthly). The "middling" months, on average, provided relatively little drama. Index A and B both outperformed the S&P 500 roughly half the time, although their average excess returns were modest in comparison to either the poor or the good months.

Based on these relationships, it seems fair to argue that Index A is likely to outperform the S&P 500 in a "poor" market environment, and underperform in a "good" market environment. Index B behaves in a contrary manner; we'd expect it to outperform in a "good" market environment and be indifferent (or worse) otherwise. This perspective helps us understand why the indices' live performance in 2011 and 2012 seems at odds with the backtest results.

Exhibit 5: Monthly Performance Distribution During the Backtest and Live Periods

	1991-2010 (%)	2011 (%)	2012 (%)
Poor	12.5	16.7	8.3
Middling	65.8	75.0	66.7
Good	21.7	8.3	25.0

Source: S&P Dow Jones Indices LLC. Data as of Dec. 31, 2012. Index performance based on total return USD. Tables are provided for illustrative purposes. Past performance is no guarantee of future results. These charts and graphs may reflect hypothetical historical performance. Please see the Performance Disclosure at the end of this document for more information regarding the inherent limitations associated with backtested performance.

Relative to our backtest period, 2011 witnessed a shortage of "good" months. There were slightly more poor months, and many more middling months, than in our backtest period.[11] Based on these results alone, we might expect that Index A's 2011 performance would be better than the backtest would seem to indicate, and that Index B's would be substantially worse. That, of course, is exactly what we observed. In 2012, the tables were turned. Now we had a relative shortage of poor months and a slight surplus of good months. Not surprisingly, Index A, which thrives in poor months, underperformed, while Index B, which tends to outperform in good months, outperformed.

Index A and Index B[12] give us a convenient way to illustrate an important principle: **the performance of many strategy indices is contingent on the market conditions under which they are measured**. In our example, Index B is generally successful in markets that trend upward. Take away the trend, or the upward direction, and Index B looks bad—but that isn't Index B's fault. Similarly, Index A is generally successful in markets that are choppy or that trend downward. Take away the chop and the downward bias, and Index A doesn't look as good, but that isn't Index A's fault either.

The critical point is that we should use history not simply to see whether a strategy index outperformed, but also to understand the market conditions, or regime, in which it does well or poorly.[13] **The nature of the market regime, and the nature of a strategy's response to that regime, are two separate things and shouldn't be conflated.** No backtest can anticipate changes in market regime, but good ones can help us understand how and to what degree a strategy's behavior is regime-dependent. The same is true for live historical data.

IS THE PAST PROLOGUE?

Our example, although it uses real indices and real data, is admittedly stylized. Parsing total return performance by market condition can be a valuable analytic exercise, but it's not a comprehensive index review. A potential user might also be interested in an index's volatility, its tracking error relative to the market, its correlation with other indices, its maximum drawdown in times of stress. And knowing an index requires more than simply knowing its returns; it can also be useful to understand what constituent securities the index holds and how rapidly they turn over.

[11] Of course the terms "slightly" and "many" need to be understood in context. No matter how we classify them, 2011 and 2012 each comprise only 12 months.
[12] Spoiler alert: Index A is the S&P 500 Low Volatility Index. Index B is the S&P 500 High Beta Index. Both indices were backtested using data through 2010, and went live on April 4, 2011.
[13] "Market conditions" can include more than just the return of the stock market. An investor might want to condition a strategy's performance on other variables—e.g., on changes in interest rates or exchange rates.

We've offered some thoughts for the potential index user who's trying to evaluate historical index performance data. It's important to repeat that, *mutatis mutandis*, **we should evaluate real historical data in the same way that we evaluate simulated data.** The same three factors—**plausibility, clarity and sensitivity to market conditions**—that help us think about simulations can also guide us in thinking about actual history. And the same attitude of skepticism is apt to serve the investor well in both cases.

ABOUT S&P DOW JONES INDICES

S&P Dow Jones Indices LLC, a subsidiary of The McGraw-Hill Companies is the world's largest, global resource for index-based concepts, data and research. Home to iconic financial market indicators, such as the S&P 500® and the Dow Jones Industrial Average℠, S&P Dow Jones Indices LLC has over 115 years of experience constructing innovative and transparent solutions that fulfill the needs of institutional and retail investors. More assets are invested in products based upon our indices than any other provider in the world. With over 830,000 indices covering a wide range of assets classes across the globe, S&P Dow Jones Indices LLC defines the way investors measure and trade the markets. To learn more about our company, please visit www.spdji.com/spindices.

PERFORMANCE DISCLOSURE

The inception date of the S&P 500 Low Volatility Index and the S&P 500 High Beta Index (the "Indices") was April 4, 2011, at the market close. All information presented prior to the index inception date is back-tested. Backtested performance is not actual performance, but is hypothetical. The backtest calculations are based on the same methodology that was in effect when the indices were officially launched. Complete index methodology details are available at www.spdji.com/spindices.

Past performance of the Indices is not an indication of future results. Prospective application of the methodology used to construct the Indices may not result in performance commensurate with the back-test returns shown. The backtest period does not necessarily correspond to the entire available history of the Index. Please refer to the methodology paper for the Indices, available at www.spdji.com or www.spindices.com for more details about the indices, including the manner in which they are rebalanced, the timing of such rebalancing, criteria for additions and deletions, as well as all index calculations. It is not possible to invest directly in an Index.

Another limitation of backtested hypothetical information is that generally the backtested calculation is prepared with the benefit of hindsight. Backtested data reflect the application of the index methodology and selection of index constituents in hindsight. No hypothetical record can completely account for the impact of financial risk in actual trading. For example, there are numerous factors related to the equities (or fixed income, or commodities) markets in general which cannot be, and have not been accounted for in the preparation of the index information set forth, all of which can affect actual performance.

The Index returns shown do not represent the results of actual trading of investable assets/securities. S&P Dow Jones Indices LLC maintains the Index and calculates the Index levels and performance shown or discussed, but does not manage actual assets. Index returns do not reflect payment of any sales charges or fees an investor may pay to purchase the securities underlying the Index or investment funds that are intended to track the performance of the Index. The imposition of these fees and charges would cause actual and back-tested performance of the securities/fund to be lower than the Index performance shown. As a simple example, if an index returned 10% on a US $100,000 investment for a 12-month period (or US$ 10,000) and an actual asset-based fee of 1.5% was imposed at the end of the period on the investment plus accrued interest (or US$ 1,650), the net return would be 8.35% (or US$ 8,350) for the year. Over 3 years, an annual 1.5% fee taken at year end with an assumed 10% return per year would result in a cumulative gross return of 33.10%, a total fee of US$ 5,375, and a cumulative net return of 27.2% (or US$ 27,200).

DISCLAIMER