Too many traders add too many rules and filters to improve trading system backtest results. There is a better way to optimize.

DON'T BUILD A BETTER BACKTEST

by Kevin J. Davey

Ask any experienced system developer about backtests, and you'll likely get an exasperated look. On one hand, he'll say, backtests are great because they can demonstrate if a trading idea has any historical merit. On the other hand, he'll counter, many times backtests tell you little or nothing about future profitability because you are curve fitting or over-fitting a system. Because of this, backtests are both a blessing and a curse.

Those new to trading, however, rarely see the duality in backtests. On the contrary, they see a world of historical profit in optimized tests, parabolic hypothetical equity curves and a sea of dollars just waiting for them in live trading. They expect the historical performance to continue well into the future. A few failed strategies later, though, the trader usually laments why the terrific backtest always fails to emerge in real-money trading.

So what is the big problem with traditional backtesting? Before we examine that, it is important to define exactly what a backtest is and what alternatives exist. These are shown in "Backtest comparison" (right), assuming testing ends on Dec. 31, 2014. First, there is the traditional backtest. This is by far the most popular test method, and also the most dangerous. Most trading software encourages this type of test. Simply pull up a chart, insert a strategy, and optimize all the parameters with all available data. The best result of the optimization is then what is traded. This method is very financially dangerous for most traders.

For many, bad experiences with a traditional backtest will lead to the next variation; backtest with an out-of-sample evaluation period. Instead of testing on the whole data history, the trader will test with the first 50% to 80% of the data, leaving the rest of the historical data untouched. The performance of the optimized system during this out-of-sample period will then be evaluated. This is a much better way when compared to traditional backtesting, although many people test and retest, which in effect converts the out-of-sample period to an in-sample period as the researcher gains familiarity with the data and bias creeps into the development process.

A step beyond out-of-sample testing is walk-forward testing. With this method, a longer out-of-sample period can be created. This approach is favored by many professional traders, although it can also become tainted through repeated testing.

A final method of testing is to simply start trading with no historical testing. This is the truest method because the test is in real time, with real money. But, it can take an extremely long time to evaluate whether the strategy is profitable. Needless to say, it also can be expensive. The traders who succeed with this method likely have well-formed trading strategies and deep understanding of a market's dynamics—based on years of experience—that allows them to "pre-qualify" a system before going live. There is no ambiguity in real-time results, though, unlike all types of hypothetical backtests.

As you can see, each of the three alternative methods of testing is more difficult than the simple "plug-and-chug" traditional backtesting method. Therefore, most people, especially newer traders, just stick with the easiest method. Traditional backtests can be very dangerous, though. Many people start to believe that improving the backtest is the goal of testing—that a better backtest is always desirable. An example trading system can easily show that is not the case.

BACKTEST OVERDRIVE

Let's call our new trader "John." John wants to develop a strategy for the gold market, using data from 2008-2012 for five years of test data. So, he pulls up a chart of gold daily bars in his trading software. He has learned from many books and websites that a moving average crossover system is a basic, and many times effective, trading system. So, he programs it into his trading software. This is what his system code looks like this:

input: mavg (2)
If Close crosses above average (close,mavg) then buy next bar at market;
If Close crosses below average(close,mavg) then sell short next bar at market.

Of course, John utilizes the optimization feature to optimize for the variable "mavg" — the moving average length. After running the optimization on 49 iterations, he gets a best net profit equity curve, shown as System A in "Optimized performance" (right). That is clearly not good enough to trade, so John embarks on a backtest improvement project.

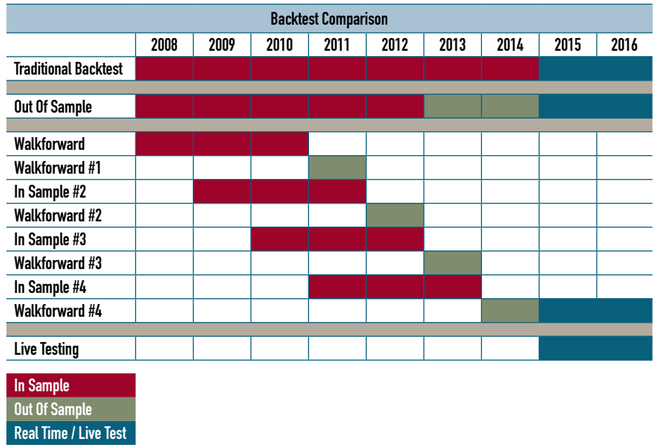
For system B, John decides that long and short markets will act differently, so the moving average lengths for long trades and short trades should be different. When he adds in this optimizable parameter, the number of iterations increases to 1,681, and his performance greatly increases (shown as System B in "Optimized performance").

Now John is feeling good about this system. But, he wants even better performance. For System C, he adds in another moving average, which he also optimizes. Now he has 8,405 iterations to optimize over. Not surprisingly, he is ecstatic when the equity curve looks much better. This is shown as System C.

However, even this performance is not enough for John. So, he adds another rule to his strategy; this time to exit after a certain number of bars. Of course, he does not know what value to use for this new rule,

BACKTEST COMPARISON

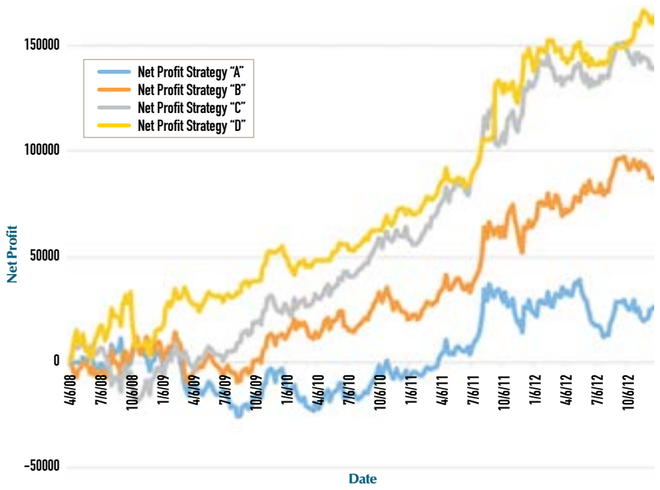
The red squares indicate the years when backtesting took place. The green squares indicate out-of-sample testing periods, or when testing was done on data that wasn't used for parameter optimization. Blue squares indicate real-money, live trading.



Source: Need source...XXXXXXXX

OPTIMIZED PERFORMANCE

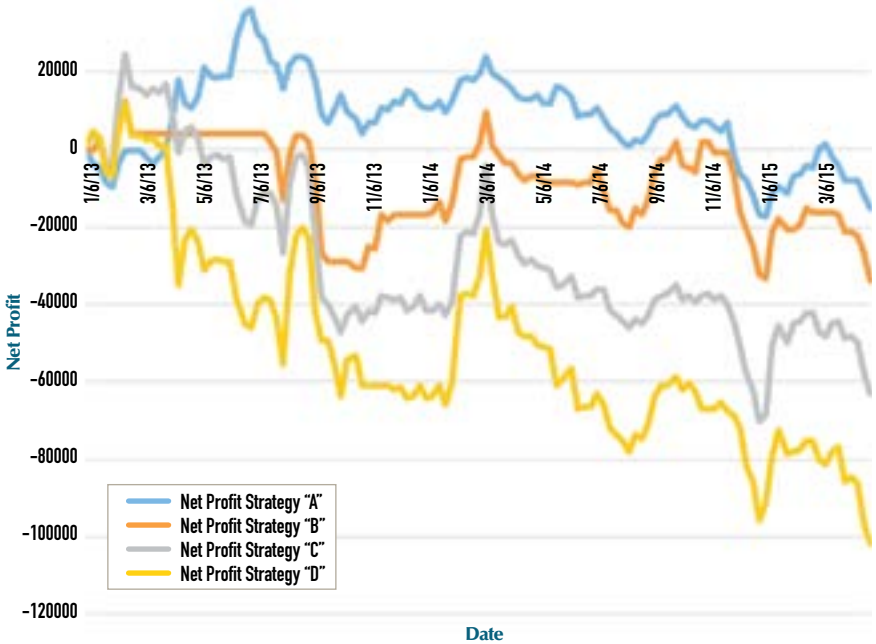
By optimizing increasingly more inputs, our hypothetical novice trader, John, produces ever-better historical equity curves.



Source: Need source...XXXXXXXX

REALITY BITES

John's optimized trading systems didn't do too well in live trading. In fact, they did terribly. This is because the systems were fit to the specifics of the in-sample data.



Source: Need source...XXXXXXXX

Real Time Net Profit				
	2013	2014	2015	Total
Strategy "A"	$11,400	-$28,370	$1,575	-$15,395
Strategy "B"	-$28,380	$13,195	-$3,455	-$18,640
Strategy "C"	-$21,090	-$16,940	$8,700	-$29,330
Strategy "D"	-$22,890	-$2,708	-$12,815	-$38,413

so he decides to optimize across what are now 19,404 iterations. Another optimization, another improvement. John now has the equity curve shown as System D.

At this point, John congratulates himself. He has turned a barely profitable moving average strategy into a historically great looking strategy. But, has he really created a better system? He obviously has generated a more impressive historical backtest, but does that mean anything? Does better historical performance translate to better real-time performance?

Unfortunately for John, and unfortunately for most people who develop strategies this way, adding rules to create a better backtest does not mean the performance in real time will be any better. In fact, many times, improving the backtest actually makes the real-time performance worse.

OUT-OF-SAMPLE REALITY

To see this, let's examine how John's strategies do in live trading. Since John's backtest was only until the end of 2012, we can examine what happened to his four strategies during 2013-15. This is shown in "Reality bites" (left). As you can see, the better-performing backtests actually have worse performance with the real-time unseen data of 2013-15. Thus, by focusing on making the backtest better, John actually made things a lot worse.

This is unfortunately a common occurrence. Many traders think they are doing the right thing by improving the backtest, when they are actually just hurting themselves. While this is not always true — sometimes adding rules to a strategy improves both the backtest and real-time performance — a trader always needs to be aware of this possibility.

These are some tips to overcome this tendency to improve the backtest:

- Set realistic expectations. Don't try to create a perfect looking equity curve. Real strategies sometimes have severe drawdowns and many flat periods. If your backtest results look too good to be true, the strategy probably will not work going forward.
- Don't keep adding rules and iterations just to improve the backtest performance. Remember, "past performance is not indicative of future results."
- Consider an alternative method of testing. Out-of-sample, walk-forward and real-money testing are all highly superior to traditional backtesting. Consider if one or more of these methods is appropriate for you.

Many traders find historical testing to be indispensable. It allows them to analyze different strategies and see which have held up over time. While it does not mean profitable performance will continue, it is reassuring to trade a method with a profitable history. The problem comes about when the trader tries too hard to create a better history. Many times, improving the backtest leads to the opposite effect in real time — worse real-time performance. Therefore, a trader always has to be careful when developing a strategy and resist the urge to build a better backtest and end up over-fitting the system. Δ

Kevin J. Davey has been trading for more than 20 years. Kevin is the author of the Wiley Finance book "Building Winning Algorithmic Trading Systems."