The Ups and Downs of Backtesting

by Ken Long

Back testing is a crucially important way of understanding a trading system and the process is an essential part of a complete trading plan. That being said, too much back testing can lead to learning the wrong lessons if you're not careful.

Traders should respect the synergy between professionalism and experience in their level of backtesting. Even an experienced trader should carefully consider the results of a detailed back test when taking on a new strategy or operating in a new time frame so that his experience provides proper expectations rather than blinds him to some new aspect of the system.

Traders can back test to various degrees depending on their experience, beliefs, and objectives. In some cases an experienced trader who is considering an idea that is similar to previously reliable systems may only need minimal back testing to be convinced that the idea is worth trading with live money at a reduced risk level. For others, the validity of a trading idea may not be convincing until they see it work over multiple time frames, in multiple markets and in all different market conditions.

The type and level of backtesting is largely a matter of personal taste; it comes down to personal choices about the risk of capital rather than a theoretical exercise in the pursuit of absolute truth. Leave that for the academics. We want to make money.

Benefits of Back Testing

Properly constructed back testing will identify whether or not an idea has a persistent edge and under what conditions it will manifest. By properly controlling for different parameters, we can isolate those which add the most value to a particular proposition. We can test for robustness and measure the sensitivity of the edge to changing parameters. From that, we may be able to identify specific market conditions where the edge is significant and tradable or identify a subset of the total market trading targets in which this idea works best.

Back testing should reveal the likely win rate percentage, the importance of slippage and commissions, the trading frequency, the maximum adverse excursion, the longest normal winning and losing streaks, and both the maximum and average figures for wins and losses.

One of the most important result sets for analysis is the distribution of R-multiple results in the form of a frequency histogram. We would like to see a somewhat normal distribution that has most of the trades clustered around the mean and with an orderly profit tail to the right. A tail of some high, positive R-multiples suggests we have the possibility of large winning trades. We would also like to see a carefully-controlled left tail of losses that suggests we are able to engineer our risk with some confidence.

Having this kind of data in hand allows us to determine where, when and under what conditions this idea is tradable and the expected results. When we proceed into live market trading the prototype system with minimal risk, we can then compare actual results to backtest results to see if the trades can be managed as intended.

Under these kinds of conditions and looking for this kind of information, back testing is an important part of the trader's repertoire.

Limitations of Backtesting

One of the primary downsides of back testing is that it can lure the trader into overconfidence about a system based on how it performed in the past.

There is always a real danger of curve fitting and data mining to find a perfect system that worked with a set of past market conditions in the past but will never occur again in the future. The market is a complex, adaptive system that never shows traders the same face twice.

Even when traders perform a backtest rigorously and with full knowledge of the limits of its ability to forecast into the future, they will commonly see a large discrepancy between backtest results and actual results from live trading.

There are several potential causes for this discrepancy. Sometimes, traders will backtest a system in isolation and not as part of a full portfolio of strategies. Often back tests may use unrealistic values for slippage and commissions. Many times the backtest will not take into account the human dimension of executing a set of trading rules. Experience shows that this is one of the most overlooked aspects in back testing.

Backtesting specifically, and any forecasting method in general, has limits and that's perhaps the biggest challenge for traders to understand. Some traders place so much reliance on their back testing that the evidence of divergent actual results will not convince them that they missed something important. They might persist in trading a system that will simply not work in the real world. (Attribute this to the overconfidence bias combined with the need to be right.)

Professional engineers and doctors are especially prone to this common problem because of their ingrained belief systems from their professions. Those professions in particular place a premium on being right to be successful, yet in the trading profession, the ability to act with incomplete knowledge and a willingness to be wrong often leads to the best results.

While back testing offers many powerful advantages, the professional trader concurrently recognizes the limits of its usefulness. The professional trader should take back testing results into consideration as a way to select a system to prototype with real money, in real markets with the human factors fully engaged to see what the real world results look like before committing to full production system risk.

An Interview with 2011 Wagner Award Winner Thomas Krawinkel: Part 1

by RJ Hixson

Thomas Krawinkel recently won the 2011 Wagner Award for his submission to NAAIM's annual research paper competition. He has been studying Van Tharp's material for several years and cited Van Tharp in the References section of his paper. We conducted this two-part interview with him after he returned from presenting his paper at the NAAIM annual conference.

What kind of trading do you do?

Well I am not really trading yet. I wanted to start trading full time a few months ago, but I ran into problems testing some swing systems last fall. My analysis of those issues and the lessons learned became the foundation of my NAAIM paper. I expect to be trading later this year on a full-time basis and am preparing for that now.

What did you do before trading?

I was a financial controller and, more recently, I was a director of a small company. I left that position last year so I could focus full time on trading. Even though I've been getting ready to trade for a living for a total of five years now, about two years ago I realized that I wanted (and needed) to spend all my time in preparation. My wife has a good position, so we decided to take a chance and allow me three years of full-time devotion to trading. I hope to last as a trader much more than three years though.

How did you find out about the NAAIM competition? What made you want to submit an entry?

I was not out looking for a competition. I heard about NAAIM in Ken Long's chat room where I was posting some research and a number of test results. One of the chat room members, had attended NAAIM's annual conference and suggested that I submit a paper to their competition. The National Association of Active Investment Managers is a US-based organization that has two to three hundred members who are traders and active investment managers.

Jerry Wagner at NAAIM started the research competition about three years ago to attract attention and new members to the organization. To make the competition attractive, he put up $10,000 for first place. That's why they call the winning prize the Wagner Award.

What was your paper about?

I had been doing a lot of back testing and found some pretty good looking systems. However, when I added in position sizing™ rules to the testing, I was stunned to see how severely the performance deteriorated. It was so bad that at first I thought I had mathematical errors in my spreadsheet. After looking into it I saw that the results were bad because the system had to skip lots of trades. Missing trades really hurt the profitability of the system.

As I looked into the root cause of why so many trades were not taken, I saw there was a limit to the buying power of my capital—sometimes there was not enough money to take all the trades because the signals were clustered together.

What were the main lessons of the paper?

When you look at a system you probably think about expectancy, variation and frequency of trades, which are the three core indicators in Van's SQN® score and expectunity. However, one factor, in my opinion, is missing: the number of times the system would try to enter multiple trades.

Every concurrent position uses a portion of capital or more accurately capital and leverage—that's what I call buying power. I found that even a system with a high SQN score (of 3.0) that signaled 600

trades spread out over 16 years had occasions when there were many signals clustered tightly together. I could have only taken all of the trades if I risked less than 0.2% of my capital per trade, which was too small of a position size for me to meet my objectives.

Overlapping trades consume buying power. If there are enough signals or already open positions all at once, your buying power (which is finite) prevents you from taking all the new signals. You have to skip trades, which is not part of your system design and therefore is not reflected in any backtesting results.

Skipping trades introduces an element of randomness into your real life performance that would not be there if you were able to execute each trade according to your rules. You simply can't control when your buying power has been used up and which trades will be skipped.

How could a trader deal with this?

1. Increase your leverage or trade leveraged vehicles, which effectively increases your buying power.
2. Shorten the duration of the trades, which gives you less overlap.
3. Reduce the amount of clustering the system experiences (signals firing at almost the same time).
4. Reduce the number of vehicles in your trading universe so you have less clustering.

If all these fail, reduce the risk per trade. These strategies, individually and as a whole, have led to my personal solution: day trade. As a result of my research, I may no longer trade swing or intermediate term systems.

What was it like to present the paper at the annual NAAIM conference?

I'm not used to giving presentations, especially in English, so it was quite exciting for me. The audience seemed to listen, and everyone said I did well.

What surprised you about the conference?

I was really surprised by how professional investment managers approach trading. If you have studied Van Tharp's material, you probably think in terms of trading signals, defining risk, measuring reward in terms of risk, and making a profit at the end of the year. They don't think along those lines. Investment managers don't mind a loss at the end of the year—as long as the market was down more than their fund. They justify their losses: "Even though I was down, I beat the market." My position as a client of such a manager, though, would be that they lost money—MY money.

Also, most of their systems seem to be rotational systems that are always invested in some sector of the market. With this method, the buying power problem does not exist: they always rebalance the equity but never have any clustered signals. They simply allocate their money where they think it will perform best. This is not the way I approach trading. If there are no signals, I'm happy to sit in cash and wait for signals. If there are many signals, I will be very active and exploit the opportunities.

Aside from the award and prize money, what did you gain from the experience?

I found most precious the appreciation of my ideas by knowledgeable people. Until last year, I didn't know any other traders and felt quite lonely. When you are working hard for so long on something, it's good to get validation that the work is valuable and interesting. It was also good to learn that I didn't get disconnected from reality in the process—especially since I have little trading experience.

Do you have any advice for others about submitting a paper for a competition?

Find any competition you can and submit something—but not for money or an award.

The biggest benefits come from sitting down and formulating your thoughts in a way that other people will be able to comprehend. Writing this paper helped me organize my own thinking in a way that would not have happened if I had not submitted it. As a result of the research and writing the paper, I have changed my approach to trading.

How so?

When I look at systems now, I'm looking for little clustering and short time horizons. As I mentioned, the solutions to the problems that I found in my backtesting of swing systems seems to point to day trading. Personally, though, I don't want to sit in front of a screen all the time; therefore, I intend to use my programming skills to make the trading process fully automated. Having talked to quite a number of traders during the last year, I now believe I am probably the most extreme mechanical trader that I know of. Maybe there just aren't many purely mechanical traders out there but that is the direction that I am heading.

An Interview with 2011 Wagner Award Winner Thomas Krawinkel: Part 2

by RJ Hixson

What are your thoughts about backtesting?

First, I want to differentiate between backtesting and optimizing. I think many people tend to confuse the two. Many people seem to think that backtesting is optimizing their rules on a set of data. This is dangerous because you can optimize each variable and get fantastic results but all you have really done is to curve fit the rules to the data sample. You will not be able to generate the same kind of results with live trading.

To me, backtesting is merely testing an idea on data. I have a set of rules that I can run on historical data. What I like to do is to pick some trades at random out of a backtest and then look at the charts for those trades.

I collect all of the trade results and analyze the expectancy, draw time lines of key indicators, apply position sizing strategies. I try to learn why each trade worked or didn't work. If I see enough of a pattern, I consider what I might change. I would never let a computer optimize or change the system for me.

My way of optimizing or changing the system comes from looking at a very small sample of trades compared to what people tend to do with testing platforms. Affordable computing power now allows people to use a platform to optimize a set of rules for all trades over a very long period of time. I think that's a recipe for disaster.

Why?

Besides curve fitting, testing problems occur in large part due to data issues. I have learned to pay a lot of attention to the data I use for backtesting.

One general problem with long term testing is survivorship bias. Only the stocks that are still around show up as part of the backtesting results, which may make the results appear to be better than what the system might actually generate.

Then there are other data problems, simple things like how the data provider handles dividends. For example, Stockcharts.com subtracts dividend payments from the stock prices prior to the payment. Google, however, does not make this adjustment.

If you look at Stockcharts.com, it appears that the price on a specific date in the past was lower than where it really traded at the exchange or as reported by Google. The dividend adjustment affects the 52-week high prices, the moving averages, and multiple other indicators. It all depends on which data provider you use.

Many people do not know about these adjustments, but they should if they base their trading system decisions on price action.

Would there be a similar adjustment for one-time events, say a special five dollar dividend?

Certain data providers would reduce all historical prices by five dollars. In some cases, these adjustments can take historical prices below zero. Others data providers will adjust the historic prices by the percentage of the dividend against the stock price at the time of payment, which in my opinion is the best way to handle that situation. This way, the price never gets below zero.

And we are talking about dividends here. Imagine the adjustments the data providers make for stock splits or reverse stock splits.

Data issues like these become extremely important when you backtest over ten or more years because dividends, splits, etc. happen regularly and affect the stock price data. People seem to think that testing more stocks over more years gives you better test results. However, more years of data can actually skew your test results rather than make them more valid.

Then why backtest at all?

Backtesting gives me a general feeling whether an idea will work. That's what I use it for. Humans tend to look for support for our opinions and ignore evidence against them. I see that in Ken Long's chat room and at the poker table. I have heard people say "I don't play Aces" even though they know that's the best hand; however, they might believe they always lose when they play Aces. They may have lost just once in ten hands playing Aces, but they remember that one loss more than the other nine wins. Backtesting helps me see that playing an Ace is a good idea—regardless of what I have heard from others—and others definitely affect your beliefs.

You can read about or hear about a trading idea that applies to a particular set of stocks at a particular time but that opens the question as to whether it works on other stocks over other time frames. If you know other traders who talk about the same idea, that can seem to make the idea more "real." But it's not real, it's just a perception that can be distorted the more you see it, hear it, or think it. The question is whether an idea is valid. That's what testing is for—to see if an idea has an edge or if it is nothing extraordinary.

When I design system rules, I get an idea of how the system works for a particular market. But when I run those rules on a different market type or in different time periods, I expect this system to generate similar results. I want to test an idea for validity and an intrinsic edge—not simply for a good curve fit.

I prefer chart patterns as a basis for trading systems because I believe they represent psychological patterns. I contend that these kinds of systems should work on data today as well as on data from 10 years ago as I don't think human psychology has changed much in the last 10 years.

Price pattern based systems work because of psychology rather than other factors—at least that's my personal belief.

I believe part of the purpose of back testing is to help me gain confidence in a system. I have seen other traders who can trade a new idea with real money right off. I don't like to do that. I like to test an idea first to see if it has worked before, rather than just in the current market.

When I was testing swing systems, I was surprised to see how widely their performance varied from year to year. One year the SQN score for a system might be 3.0, another year it was 5.0. Another

year the system might have had a negative expectancy. I think it's important to see the size of the swings. Looking at a single number over the backtest period as a whole, even the SQN score, won't tell you if you could actually trade the system and stomach the variation over different periods. My conclusion from this is that people who want to trade swing systems need to be prepared for long drawdowns.

Most people look at win rate, average R-multiple, accumulated earnings, and results like that. I think you've got to be able to take a large set of data and split that into different chunks and examine each chunk separately. How do aspects of the system change from chunk to chunk?

Did you look at the market types for these chunks?

I did; however, I didn't find much correlation between system performance and market types. Although for example, a certain swing system might get a lot fewer signals in a bear market than a bull market. The way I was testing may have affected the correlation results but the differences were very low—in the range of statistical noise for me.

What's your advice for others doing backtesting?

If you want to achieve the kind of live trading results you saw in your backtesting results, you will have to trade a system the exact same way it was run in the tests. Here are the questions you need to ask:

- Can it be done?
- Can I do it? Is it possible with my execution capabilities (i.e., automated vs. manual), my psychological setup and my buying power?

If you can't actually trade the system the way the test was run, revise the test to reflect the way you can actually trade it.

I would also add that if I'm sick or on holiday and not trading, when I return, I simulate the trades that I missed while I was out. I will enter those simulated trades in my trading log in order to help me generate a full sample for an accurate comparison to the backtested results.

Missing trades (anything more than even just 5% of the trades) can really skew your system results. I was very surprised at how big of an effect missing just a few trades can have on system results whether because of a cluster problem as I wrote about in the paper, you're on vacation, or whatever else causes you to miss those trades.

When I research trading systems now, I look for little clustering and short time horizons—I don't try other system "adjustments." For example, I've heard people say "Let me increase the 'expectunity' of the system by trading a larger pool of instruments." They think that they can go from trading 10 ETFs to trading all 500 companies in the S&P. Unfortunately for them, the larger pool increases the problem of clustering: they won't have the buying power to take all of the system's trades. All they do then is add randomness that may take their real life performance far from what they would expect to see by following a tested system.

I want to get back to the point about looking at your test and understanding what you can actually trade—you really have to focus on that. You may want to make $50,000 this year and the key performance indicators of a system derived from all trades in a test may suggest that you can do that. But you have to determine whether your buying power will cause you to skip a significant number of trades. I designed a Quick Check spreadsheet that is available here. Through it you may find that your "$50,000" system will only make you $10,000 given your personal circumstances. It would be better to know you have a $10,000 system before you quit your job or count on that trading income in any way.