# Market Fox

'The fox knows many things…'

**APRIL 3, 2018**
**MARKET FOX**

# A Better Back Test Checklist for Non-Quants

I recently had the pleasure of interviewing James O'Shaughnessy (@jposhaughnessy (https://twitter.com/jposhaughnessy)) for an upcoming episode of the i3 Podcast (https://i3-invest.com/category/podcasts/). Jim is the chairman and CIO of O'Shaughnessy Asset Management (http://www.osam.com/). He's also the author of What Works on Wall Street, The Classic Guide to the Best-Performing Investment Strategies of All Time (https://www.amazon.com/What-Works-Wall-Street-Fourth/dp/0071625763/ref=sr_1_1?ie=UTF8&qid=1522709878&sr=8-1&keywords=what+works+on+wall+street+james+oshaughnessy), now in its 4th edition. *What Works on Wall Street* is one of **the** best books on applied quantitative investing. The analysis and results are presented clearly and simply. Even a non-quant such as me can understand and apply the material.

During our conversation, I mentioned that I wanted to put together a checklist for non-quants to help them identify poorly constructed or misleading back tests and spurious quantitative results. Jim was kind enough to share a post that he'd written for his blog (http://jimoshaughnessy.tumblr.com/post/98397869279/the-power-of-back-testing-investment-strategies) on what makes a good back test. It has some very helpful suggestions. For example, Jim explains why its important to randomly resample data to reduce the risk of data mining:

> *Another technique that we employ is bootstrapping the data. Bootstrapping randomly resamples the overall results for the various strategies we test obtained by running 100 randomly selected subperiods to make certain that none of the randomly selected periods vary to any significant degree from the overall results shown for the various strategies. Typically, we view a factor as useful or predictive when there is a large spread between the annualized returns of the best and worst decile of that factor. The fact that the best decile of stocks with the best (highest) six-month price momentum beats the worst decile (stocks with the worst price momentum) by 9.96 percent per year for the last 83 years is powerful information that greatly influences how we advocate managing money. To eliminate any potential sample bias in this analysis we run a test on randomly selected sub-samples of the data to make sure that similar decile return spreads exist regardless of the group of stocks that we are considering. For each of the 100 iterations of each bootstrap test, we first randomly select 50 percent of the possible monthly dates in our backtest and discard the other 50 percent. We then randomly select 50 percent of the stocks available on each of those dates and discard the rest. This gives us just 25 percent of our original universe on which to run our decile analysis. We do this 100 times for each factor and analyze the decile return spreads. It so happens that for our best factors, the return spread between the best and the worst decile remain consistent in these 100 iterations. Said another way, for the six-month price appreciation factor no matter which group of stocks are possible investments, it is always better to buy the decile with the best price momentum. If we discovered that there were large inconsistencies in the bootstrapped data, we would have less confidence in the results and investigate if there was any evidence of unintentional data mining inherent in the test.*

Many back tests consist of the application of one or more decision rules on a sample of historical data. But this ignores the fact that history, as we know it, is only one of several possible and plausible sequences of events that could have happened. Resampling attempts to get around this problem by using the data to create many different alternate histories. We can be more confident that that if a result survives resampling, it's less likely to be a one-off.

**Jim's post was a big help in putting together my checklist. I'd also like to thank the following people whose advice and suggestions helped me in putting together a better back test checklist for non-quants**:

- Troy Rieck, Executive Officer, Investment Strategy at Equip (@RhinoTroy (https://twitter.com/RhinoTroy))
- Corey Hoffstein, Co-Founder and Chief Investment Officer at Newfound Research (@choffstein (https://twitter.com/choffstein))
- Professor Jack Gray, Adjunct Professor of Economics, UTS Business School

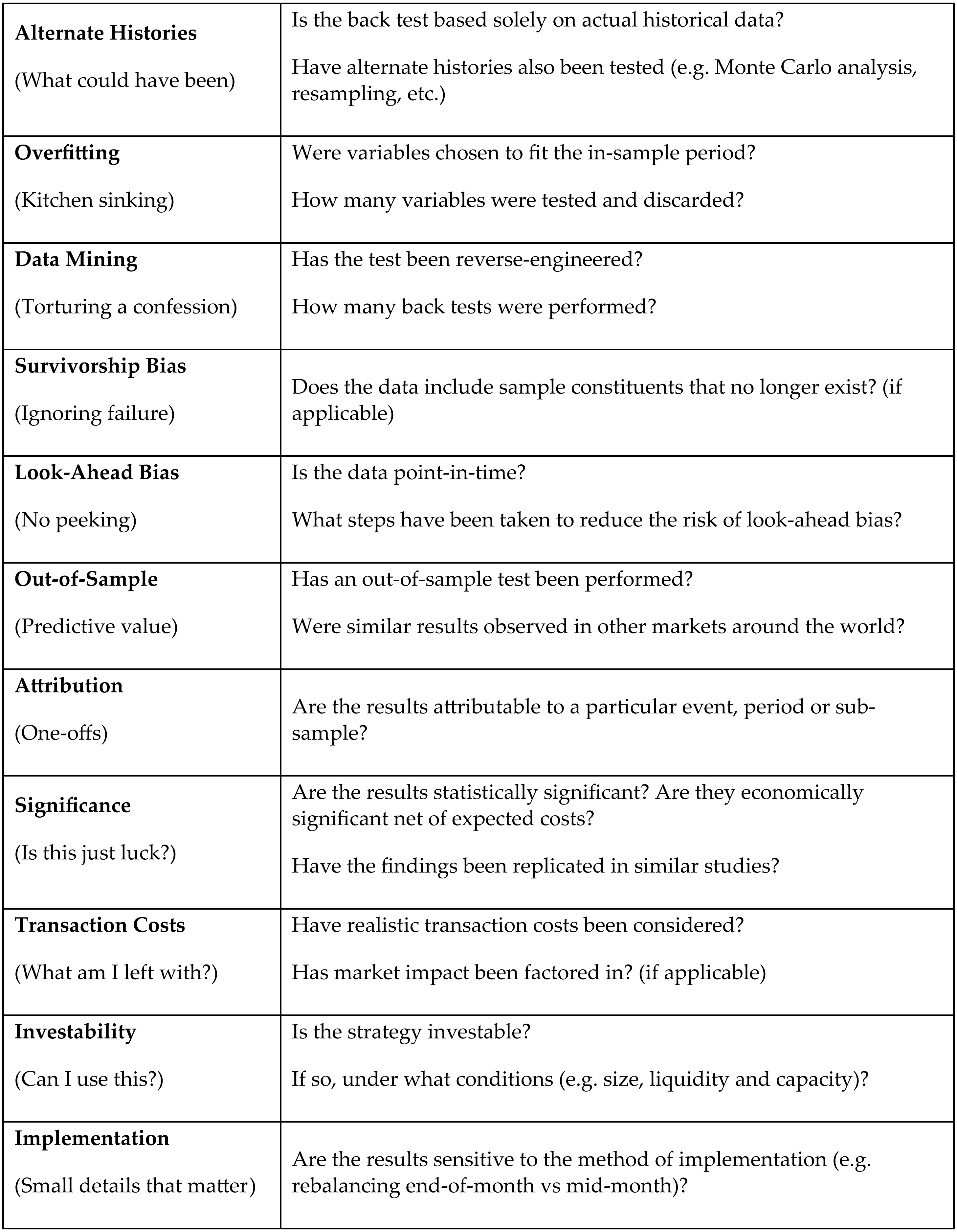
As it turns out, Corey has written an excellent paper entitled Backtesting: Problems & Solutions (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2209310) which was also very helpful. Corey and the Newfound team publish a lot of high-quality quantitative research on their research blog Flirting with Models. (https://blog.thinknewfound.com/)

Jack will soon feature in episode of the i3 Podcast (https://i3-invest.com/category/podcasts/). He had me laughing for most of our conversation. Jack has a PhD in mathematics and a career working in quantitative finance with firms such as AMP in Sydney and GMO in Boston. We explore the reasons why Jack is sceptical about a lot of quantitative finance research. It's an episode that you won't want to miss.

# The Checklist

**The checklist's objective is to help users determine how much confidence to place in a back test or a set of statistical results.** No back test is perfect. Every back test suffers from limitations and the possibility of errors and biases affecting results. While these problems can't be avoided, they can be minimised and managed. **Our task as users of back tests is to find a back test whose limitations we can live with**. In other words, the back test's limitations, degree of bias or chance of error are not so bad that they completely invalidate the results.

| Criteria | Question |
|---|---|
| **Rationale**<br>(Correlation ≠ Causation) | Is there a sound economic rationale (e.g. a theory of causal relationship) supporting the indicator or strategy being tested? |
| **Method**<br>(Replicability) | Have the test's assumptions, rules, constraints, data, etc. been clearly outlined?<br>Has sufficient detail been provided to allow an independent party to replicate the back test?<br>Have you engaged an independent, reputable researcher to try and reproduce your results? |
| **Type I and Type II Errors**<br>(Drawing wrong conclusions) | What steps have you taken to minimise the chance of Type I (false positive) and Type II (false negative) errors? |
| **Data**<br>(Reliability) | Is the data accurate and clean?<br>Has this been verified? |
| **Sample Size**<br>(Enough data?) | Is the sample size sufficiently large?<br>Is there enough historical data? |
| **Representativeness**<br>(Realistic or make believe?) | Are the test rules realistic?<br>Do they accurately represent how an investor with similar risk/return objectives would invest? |
| **Benchmark**<br>(Opportunity cost) | Is the chosen benchmark appropriate?<br>Is the benchmark investable?<br>Is the back test strategy allowed to hold off-benchmark bets? If so, why? And how much? |

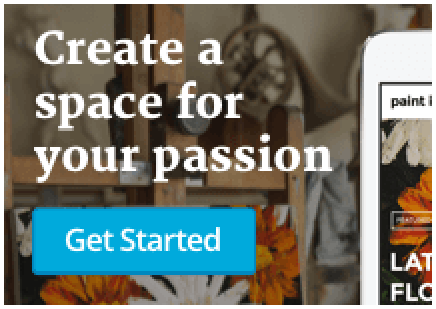


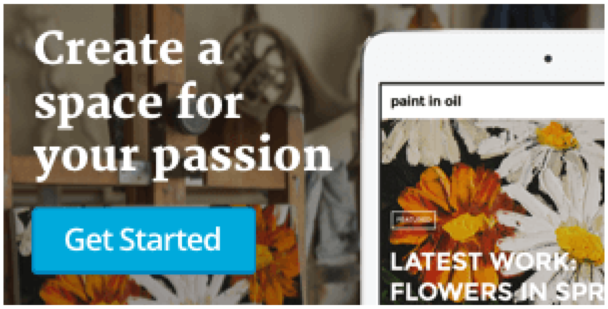
| | |
|---|---|
| **Alternate Histories**<br>(What could have been) | Is the back test based solely on actual historical data?<br>Have alternate histories also been tested (e.g. Monte Carlo analysis, resampling, etc.) |
| **Overfitting**<br>(Kitchen sinking) | Were variables chosen to fit the in-sample period?<br>How many variables were tested and discarded? |
| **Data Mining**<br>(Torturing a confession) | Has the test been reverse-engineered?<br>How many back tests were performed? |
| **Survivorship Bias**<br>(Ignoring failure) | Does the data include sample constituents that no longer exist? (if applicable) |
| **Look-Ahead Bias**<br>(No peeking) | Is the data point-in-time?<br>What steps have been taken to reduce the risk of look-ahead bias? |
| **Out-of-Sample**<br>(Predictive value) | Has an out-of-sample test been performed?<br>Were similar results observed in other markets around the world? |
| **Attribution**<br>(One-offs) | Are the results attributable to a particular event, period or sub-sample? |
| **Significance**<br>(Is this just luck?) | Are the results statistically significant? Are they economically significant net of expected costs?<br>Have the findings been replicated in similar studies? |
| **Transaction Costs**<br>(What am I left with?) | Have realistic transaction costs been considered?<br>Has market impact been factored in? (if applicable) |
| **Investability**<br>(Can I use this?) | Is the strategy investable?<br>If so, under what conditions (e.g. size, liquidity and capacity)? |
| **Implementation**<br>(Small details that matter) | Are the results sensitive to the method of implementation (e.g. rebalancing end-of-month vs mid-month)? |

| | |
|---|---|
| **Robustness**<br>(Occam's Razor) | Is the strategy as simple as possible but no simpler? |
| **Change**<br>(The 15-year old Kentucky Derby winner) | Have you considered what may have changed during the back-test period (e.g. brokerage costs, regulation, taxes, etc.) and how this might affect results? |
| **Honesty**<br>(No answer IS an answer) | What aspects of the study design and back test caused you the most concern during your research? |

I've deliberately avoided adding a score to the checklist. The idea is to think carefully about potential problems and if they've been dealt with in a satisfactory way. It's not just a check-a-box exercise.

We can't avoid back tests. They are ubiquitous in finance and investing. But we can create tools such as checklists to help us use them sensibly and safely.





DECISION MAKING EVIDENCE BASED INVESTING
INVESTMENT STRATEGY QUANTITATIVE

#INVESTING #STOCKS QUANTITATIVE

Published by Market Fox

*Student of all things investment, writer and blogger. Day job: Portfolio Manager specializing in asset allocation and multi-manager portfolio construction. View all posts by Market Fox*