How Often Does the Evaluator Promote a Strategy With No Edge?

Will Sackfield, Open Family Office · https://open-familyoffice.com/research/evaluator-calibration

Abstract

A rule that decides which trading strategies deserve real money must reject strategies with no edge and promote strategies with one. We calibrate the decision rule of this project's evaluator on synthetic strategies whose edge is known. A strategy becomes a candidate when its deflated Sharpe ratio is at least 0.95. A candidate is paper traded when a studentized stepwise test rejects it against its benchmark at a 10% family-wise error rate, and its excess return is positive. Over 200 replicates per setting, the rule paper-traded between 1% and 7.5% of the best of 1 to 50 trials that had no edge. With heavy-tailed, autocorrelated active returns, it paper-traded 3% to 9%. The first stage is much weaker: up to 39% of edgeless trials became candidates after ten years, because the deflated Sharpe ratio tests a strategy against cash, and the benchmark's own return carries it. The rule is also slow to find a real edge. An annual information ratio of 1 was found in 52% of replicates after 10 years of daily data, and an information ratio of 0.5 in 12%. A trial that held the benchmark at 1.5 times, plus active noise with no edge, was paper-traded in 13.5% of replicates after 3 years and 29.5% after 10. A Ledoit–Wolf test of the Sharpe ratio difference rejected that trial in only 10.5% to 11% of replicates. In this synthetic world, the rule held false promotion near its level, but it rewarded holding more of the benchmark, and its first stage did not separate skill from market exposure. Extensions with 1,000 replicates per setting changed little of this. Costs lowered false promotion to 1% to 3.3%, and the power for an information ratio of 1 to 31.5%. The same StepM on each trial's alpha cut the levered trial's rate from 35.3% to 6.3% at equal power. Counting correlated trials as fewer effective trials raised false candidates to as much as 79%.

Status: Published 2026-11-28. Changes since are in the addenda at the end.

1 Introduction

Many backtested strategies fail out of sample. On 888 algorithms, backtest Sharpe ratios barely predicted their out-of-sample Sharpe ratios, and the gap grew, weakly, with the total days of backtesting that each author ran (Wiecki et al., 2016). Published equity predictors lost about 58% of their return after publication (McLean & Pontiff, 2016). A sound evaluator must therefore account for the search that produced a strategy, as well as for its noise.

The literature gives the parts of such an evaluator. The probabilistic and deflated Sharpe ratios adjust a Sharpe ratio for its sample length, its higher moments and the number of trials (Bailey & López de Prado, 2012, 2014). Romano and Wolf's stepwise method names which of many strategies beat a benchmark, with the family-wise error controlled (Romano & Wolf, 2005). Yet neither the probabilistic nor the deflated Sharpe ratio has had its false-positive rate or its power measured on simulated strategies with a known edge. A rule built from these parts has not been calibrated as a whole either, applied to the best of a family of trials, in the setting where a project will use it.

This paper measures that behaviour for the decision rule of this project. It answers the project's first research question, how strategies and portfolios should be evaluated (`docs/research-questions.md`, RQ1), for the rule as it stands. The contributions are these:

- A calibration of the rule's two stages on synthetic strategies with a known edge or none, with the rates of false promotion and of detection, and their 95% intervals.
- A measurement of the rule's power as the length of the record grows, for annual information ratios of 0.5, 1 and 2, with the share of found edges that each stage stops.
- Two weaknesses that the parts' own properties do not show: the first stage promotes market exposure, and the second stage rewards holding more of the benchmark, which a Sharpe-ratio test would not.

2 Related work

Bailey and López de Prado (2012) give the probabilistic Sharpe ratio, the probability that the true Sharpe ratio exceeds a threshold that the user chooses. Bailey and López de Prado (2014) deflate the best Sharpe ratio of many trials by setting that threshold to the expected maximum Sharpe ratio of the trials with no skill. Neither compares a strategy with a benchmark, so a strategy that holds the market inherits the market's Sharpe ratio. Lo (2002) gives the standard error of a Sharpe ratio under serial correlation. The deflated Sharpe ratio counts independent trials, and correlated trials count as fewer of them (López de Prado, 2019).

Romano and Wolf (2005) control the family-wise error of many comparisons with a benchmark by a studentized bootstrap. Their statistic may be a difference of mean returns, a difference of Sharpe ratios or an alpha. Ledoit and Wolf (2008) test a difference of two Sharpe ratios with a studentized block bootstrap. Hansen's test of superior predictive ability asks the related question of whether any trial beats the benchmark (Hansen, 2005). The rule uses StepM on differences of returns, and this paper measures what that choice costs when a strategy holds more of its benchmark.

Harvey and Liu (2020) argue that multiple-testing methods must be judged by their missed discoveries as well as their false ones, and estimate both rates by a double bootstrap. This paper measures both rates too, on synthetic data where the truth is known. The probability of backtest overfitting is another guard against selection (Bailey et al., 2017), which the rule reports without acting on it. DeMiguel et al. (2009) show that the 1/N portfolio is hard to beat, which is why the rule compares each strategy with a passive benchmark at all. Statistics cannot detect leaked inputs (Gençay, 2026), which is why the rule also requires a planted look-ahead oracle to be caught.

3 Method

The rule. The rule is version 1 of the project's decision rule. It judges the selected trial of a family of N trials, all observed on the same T days. Let $r_{k,t}$ be trial k's return on day t, and b_t the benchmark's return. The rule computes the deflated Sharpe ratio (DSR) of the selected trial: the probability that its true Sharpe ratio exceeds the expected best Sharpe ratio of N trials with no skill (Bailey & López de Prado, 2014). The rule takes N as the number of trials and the variance of their Sharpe ratios as given. With one trial, the threshold is zero, and the DSR is the probabilistic Sharpe ratio against zero. The rule has three outcomes.

- **Rejected:** the DSR is below 0.95.
- **Candidate:** the DSR is 0.95 or more.
- **Paper trade:** a candidate that the studentized StepM rejects against the benchmark, with a positive mean excess return.

StepM tests the differences $d_{k,t} = r_{k,t} - b_t$ of every trial in the family at a family-wise error rate of 10% (Romano & Wolf, 2005). It studentizes each mean with a Newey–West error of 5 lags, and draws 199 circular block bootstrap samples with blocks of 5 days. A paper trade is a trial that the project would trade with paper money, and so the strongest verdict the rule gives on a backtest. The rule also requires that a planted look-ahead oracle is caught, which we assume. It reports the probability of backtest overfitting without acting on it, which we leave out.

The synthetic market. Each replicate draws T daily returns, with 252 days a year and no costs, tax or risk-free rate. The benchmark's return is

$$b_t \sim \mathcal{N}(0.0003,\ 0.01^2),$$

an annual Sharpe ratio of about 0.48. Each trial holds the benchmark and adds an active return:

$$r_{k,t} = m\, b_t + a_{k,t}, \qquad a_{k,t} = \mu_k + 0.005\, \varepsilon_{k,t},$$

where m is the multiple of the benchmark held, 1 unless stated, and $\varepsilon_{k,t}$ is a standard normal shock. The active return's standard deviation is 0.005 a day, about 8% a year. A trial's annual information ratio is $\mathrm{IR}_k = \mu_k \sqrt{252}\ / 0.005$. A trial with $\mu_k = 0$ has no edge. Because every trial holds the same benchmark, any two trials correlate at about 0.8.

Selection. In each replicate, the trial with the highest Sharpe ratio is selected, as a researcher would select it. The family's other trials enter the DSR and StepM. A trial with an edge is "found" when it is the selected trial and the rule paper-trades it.

Replicates. Each setting has 200 replicates, each with its own seed, set in advance. A rate's 95% interval is the Wilson interval. Appendix A gives the commands.

The figures are drawn from the recorded results. Figure 2 moves: it adds the years of data one at a time, because its point is how detection grows with the record.

4 How often does the rule promote a trial with no edge?

Setup. Every trial has $\mu_k = 0$. The family has N of 1, 10 or 50 trials, over 3 or 10 years.

Result. The rule paper-traded between 1% and 7.5% of the selected trials (Figure 1, Table 1). StepM's 10% bounds the chance of any false rejection in the whole family, and the DSR gate filters the trials first, so these rates are a conservative view of StepM's size. The highest interval reaches 12%. The first stage is much weaker. Between 16% and 21.5% of the selected trials became candidates over 3 years, and between 35.5% and 39% over 10 years. The candidate rate did not fall clearly as N grew.

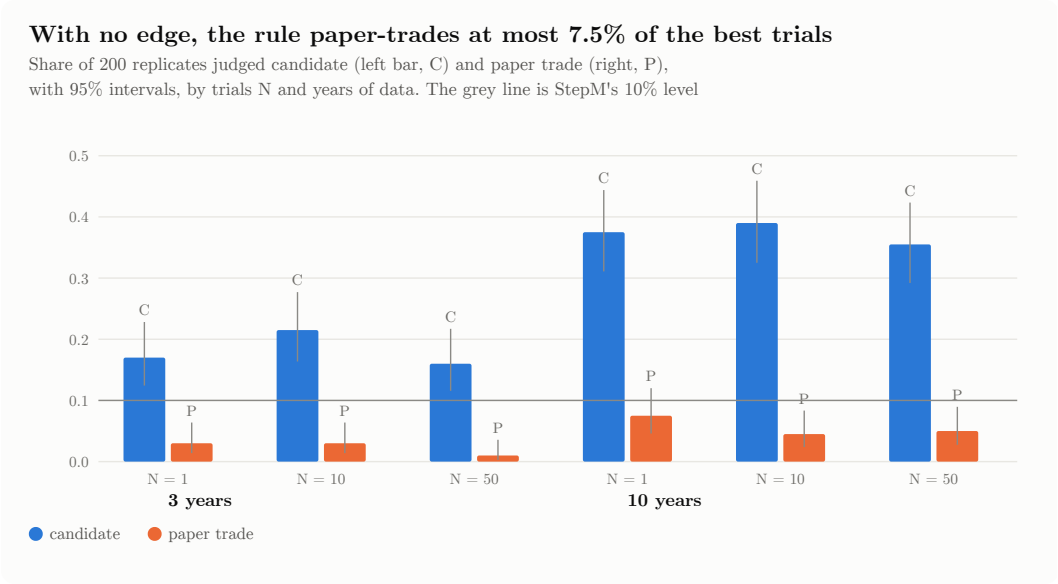


Figure 1. The share of 200 replicates in which the selected trial, with no edge, was judged a candidate (left bar) or a paper trade (right bar), with 95% intervals. The grey line is StepM's 10% level.

Table 1. The rates under the null, with 95% intervals.

Years	N	Candidate	Paper trade	Median DSR
3	1	17.0% (12.4–22.8)	3.0% (1.4–6.4)	0.789
3	10	21.5% (16.4–27.7)	3.0% (1.4–6.4)	0.783
3	50	16.0% (11.6–21.7)	1.0% (0.3–3.6)	0.770
10	1	37.5% (31.1–44.4)	7.5% (4.6–12.0)	0.917
10	10	39.0% (32.5–45.9)	4.5% (2.4–8.3)	0.917
10	50	35.5% (29.2–42.4)	5.0% (2.7–9.0)	0.915

Reading. The DSR asks whether a trial's Sharpe ratio exceeds what luck would give, not whether it exceeds the benchmark's. Every trial here holds the market, whose Sharpe ratio is about 0.48, so a long record makes the market's own return significant. Adding active noise lowers a trial's Sharpe ratio to about 0.43. Over 10 years that gives a statistic near $0.43\sqrt{10} \approx 1.35$, whose normal probability is about 0.91, close to the observed median DSR of 0.917. The minimum track record length of Bailey and López de Prado (2012) makes the same point. A Sharpe ratio of 0.43 needs about 3,700 days, or 14.6 years, to pass at 95% against zero, so the median edgeless trial would become a candidate soon after. The deflation by N did little here, which is consistent with the trials' high correlation: ten trials that correlate at 0.8 count as only about 2.8 independent ones

(Bailey & López de Prado, 2014). The second stage removes most of the candidates, because it compares the trial with the benchmark.

5 How often does the rule find a real edge?

Setup. One trial in a family of 10 has an annual information ratio of 0.5, 1 or 2, and nine have none. The record runs for 1, 2, 3, 5 or 10 years.

Result. Detection grows slowly with the record (Figure 2, Table 2). An information ratio of 2 was found in 31.5% of replicates after 3 years and 90.5% after 10. An information ratio of 1 was found in 20.5% after 5 years and 52% after 10. An information ratio of 0.5 was found in at most 12% of replicates, after 10 years.

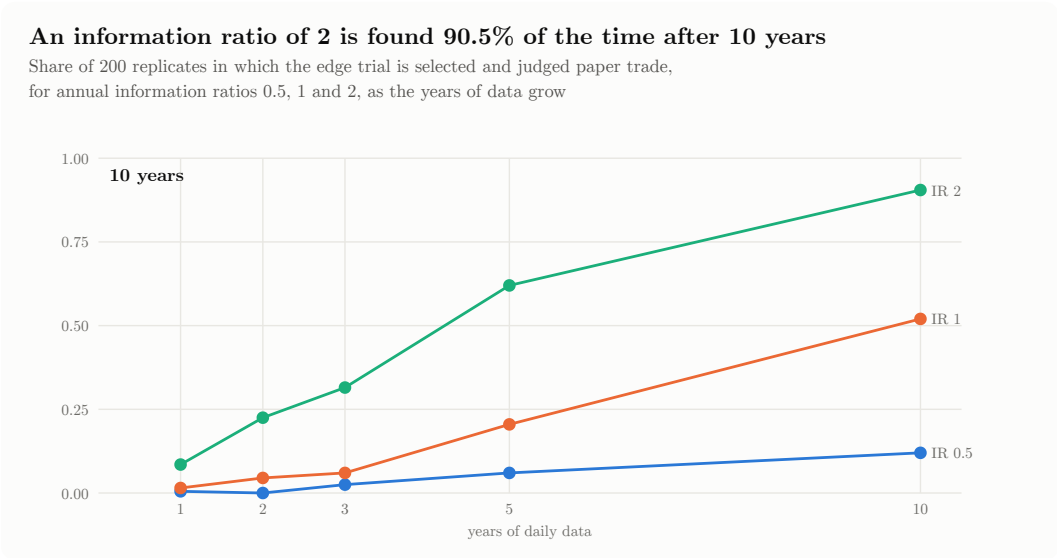


Figure 2. The share of 200 replicates in which the one trial with an edge was selected and paper-traded, against the years of daily data, for annual information ratios of 0.5, 1 and 2. The years appear one at a time, and the last frame holds every year.

Table 2. The share of replicates in which the edge trial was selected and became a candidate, and the share in which it was then paper-traded (found).

Information ratio	Stage	1 year	2 years	3 years	5 years	10 years
0.5	candidate	1.0%	2.0%	5.5%	12.5%	26.5%
0.5	found	0.5%	0.0%	2.5%	6.0%	12.0%
1	candidate	4.0%	9.5%	13.5%	26.0%	57.5%
1	found	1.5%	4.5%	6.0%	20.5%	52.0%
2	candidate	13.0%	26.0%	33.0%	62.0%	90.5%
2	found	8.5%	22.5%	31.5%	62.0%	90.5%

Reading. Three hurdles stand between an edge and its discovery. The edge trial must be selected, which fails when a lucky null trial has a higher Sharpe ratio. It must pass the DSR, and then StepM. For an information ratio of 1 over 3 years, StepM stopped more than half of the edge trials that had passed the DSR (13.5% against 6%). For an information ratio of 2, it stopped almost none once the record reached 5 years. For an information ratio of 0.5, the rates over the first three years are too low to tell apart with

200 replicates. A record of a few years can therefore tell a strong edge from noise, but rarely a modest one.

6 Does the rule reward holding more of the benchmark?

Setup. A single trial holds the benchmark at 1.5 times, with active noise but no edge: $m = 1.5$ and $\mu = 0$. Its difference from the benchmark, $d_t = 0.5\, b_t + a_t$, has a mean of half the benchmark's mean and an annual Sharpe ratio of about 0.34. The trial's own Sharpe ratio, about 0.45, is a little below the benchmark's 0.48. We also run the Ledoit–Wolf test of the difference of the two Sharpe ratios on the same trial (Ledoit & Wolf, 2008), with the same bootstrap settings, and record how often it rejects at 10%.

Result. The rule paper-traded the trial in 13.5% of replicates over 3 years and 29.5% over 10 years (Table 3). The Ledoit–Wolf test rejected the difference of Sharpe ratios in 11% and 10.5% of replicates, close to its 10% level.

Table 3. The rates for a trial that holds the benchmark at 1.5 times, with 95% intervals, and the rate at which the Ledoit–Wolf test rejects at 10%.

Years	Candidate	Paper trade	Median DSR	Ledoit–Wolf rejects
3	16.0% (11.6–21.7)	13.5% (9.4–18.9)	0.785	11.0%
10	35.5% (29.2–42.4)	29.5% (23.6–36.2)	0.917	10.5%

Reading. StepM tests the mean of the differences $r_t - b_t$, which here is positive because the trial holds more of a market with a positive premium. Over 10 years, that mean is significant often enough to pass at 10%. The rule's test is therefore a test of excess return, not of risk-adjusted excess return. A trial that needs no skill passes it as soon as the market's premium is large enough. The Ledoit–Wolf test compares the two Sharpe ratios instead, and its rejections stayed near its level, because holding more of the benchmark does not raise a Sharpe ratio when there is no risk-free rate.

7 Does the control hold with heavy tails and autocorrelation?

Setup. The null of Figure 1, with N of 1 or 10, but each active shock is drawn from a Student's t distribution with 3 degrees of freedom, scaled to unit variance, and follows an AR(1) process with coefficient 0.2.

Result. The rule paper-traded between 3% and 9% of the selected trials (Table 4). The highest interval reaches 13.8%, for one trial over 10 years. The candidate rates, between 19.5% and 42.5%, are close to those under normal shocks.

Table 4. The rates under the null with heavy-tailed, autocorrelated active returns, with 95% intervals.

Years	N	Candidate	Paper trade
3	1	19.5% (14.6–25.5)	5.0% (2.7–9.0)
3	10	22.5% (17.3–28.8)	3.0% (1.4–6.4)
10	1	42.5% (35.9–49.4)	9.0% (5.8–13.8)
10	10	41.5% (34.9–48.4)	7.5% (4.6–12.0)

Reading. The rates are a little higher than under normal shocks, but the intervals overlap, so 200 replicates cannot separate the two. This is consistent with Romano and Wolf's own simulations, in which the studentized StepM held or undershot its level under AR(1) returns (Romano & Wolf, 2005). They did not test heavy tails, and this paper does not test which part of the method holds the level.

8 Extensions: costs, PBO, correlated trials, an alpha test and a harder market

The first run left out five things that could change its rates (#307). This section adds each one. Every new cell runs 1,000 replicates instead of 200, so a 95% interval is about 1.5 to 3 points wide on each side. The first 27 cells keep their seeds, and their results are unchanged.

Costs. Each trial now pays the project's IBKR cost model (`open_familyoffice/costs.py`) every day. It trades 10% of its capital at a quoted spread of 10 basis points. The trial that holds 1.5 times the benchmark also borrows the extra 0.5 at 5% a year. The benchmark pays nothing, as buy-and-hold would. Table 5 gives the rates.

Table 5. The rates with costs, over 1,000 replicates, with 95% intervals, beside the first run's rates without costs.

Setting	Candidate	Paper trade	Without costs (200 replicates)
No edge, 1 trial, 3 years	15.8% [13.7, 18.2]	3.3% [2.4, 4.6]	17%, 3%
No edge, 10 trials, 3 years	13.9% [11.9, 16.2]	1.1% [0.6, 2.0]	21.5%, 3%
No edge, 1 trial, 10 years	28.9% [26.2, 31.8]	2.3% [1.5, 3.4]	37.5%, 7.5%
No edge, 10 trials, 10 years	29.5% [26.8, 32.4]	1.0% [0.5, 1.8]	39%, 4.5%
Edge IR 1, 10 trials, 10 years	46.2% [43.1, 49.3]	31.8% (found 31.5% [28.7, 34.5])	59%, 52.5% (found 52%)
1.5× benchmark, 10 years	25.9% [23.3, 28.7]	12.5% [10.6, 14.7]	35.5%, 29.5%

Costs lowered every rate. False promotion fell to between 1% and 3.3%. The levered trial still passed the rule in 12.5% of replicates after ten years, even after it paid for its borrowing. Costs also cut the power: an information ratio of 1 was found in 31.5% of replicates over ten years, down from 52%.

PBO. We computed the probability of backtest overfitting (Bailey et al., 2017) with 8 blocks, so 70 splits, on each family. With no edge, PBO's median was 0.49 to 0.5. A level of 0.5 would flag 46.6% to 49.6% of these families, and a tenth of them had PBO below 0.19 to 0.26. With an information ratio of 1 over ten years, PBO's median was 0.10, nine in ten families were below 0.39, and 5.5% were above 0.5. With an information ratio of 2, PBO was 0 in at least nine families out of ten. So a level near 0.2 would keep 90% of edgeless families out and let at least half of the IR 1 families through. PBO on its own is a weak filter for an edge of IR 1 or less, but it separates edgeless families from strong edges well.

Correlated trials. The trials share the benchmark, so the DSR's trial count could be too high (López de Prado, 2019). We counted the effective trials of each edgeless family in three ways. The Marchenko–Pastur count of eigenvalues above noise gave a median of 1 trial. The clustering of the project's `independence` module gave 2. The Marchenko–Pastur count on the active returns, which are what the selection chooses between, gave 0. Each count, put into the DSR in place of N, raised false candidates sharply (Table 6).

Table 6. The candidate rate of edgeless families when the DSR counts the trials in three ways, with 95% intervals.

Edgeless family	DSR with N	With the eigenvalue count	With the clusters
10 trials, 3 years	17.4% [15.2, 19.9]	41.4% [38.4, 44.5]	30.3% [27.5, 33.2]
50 trials, 3 years	15.9% [13.8, 18.3]	53.8% [50.7, 56.9]	43.1% [40.1, 46.2]
10 trials, 10 years	40.5% [37.5, 43.6]	66.1% [63.1, 69.0]	55.0% [51.9, 58.1]
50 trials, 10 years	38.4% [35.4, 41.5]	78.7% [76.1, 81.1]	70.4% [67.5, 73.2]

The correlation here comes from a shared market exposure, not from shared selection. The active returns are independent, so the search really did try N independent bets. The estimators count common factors, not independent bets, and with one common factor they count one. In this world, the DSR's raw count of trials is the right one, and the effective-count estimators would weaken the first stage further.

An alpha test inside StepM. We ran StepM on each trial's alpha, $r_{k,t} - \widehat{\beta}_k b_t$, tested against zero, beside the rule and without changing it. A trial passed this variant when it was a candidate, StepM rejected its alpha, and its mean alpha was positive. Table 7 compares the two.

Table 7. The paper-trade rate of the rule and of the alpha variant, over 1,000 replicates, with 95% intervals.

Setting	The rule's paper trade	With the alpha test
No edge, 1 trial, 3 years	4.2% [3.1, 5.6]	4.6% [3.5, 6.1]
No edge, 10 trials, 3 years	2.3% [1.5, 3.4]	2.7% [1.9, 3.9]
No edge, 1 trial, 10 years	5.9% [4.6, 7.5]	6.0% [4.7, 7.7]
No edge, 10 trials, 10 years	4.4% [3.3, 5.9]	4.6% [3.5, 6.1]
1.5 × benchmark, 3 years	14.7% [12.6, 17.0]	3.5% [2.5, 4.8]
1.5 × benchmark, 10 years	35.3% [32.4, 38.3]	6.3% [5.0, 8.0]
Edge IR 1, 10 trials, 10 years	51.2% [48.1, 54.3]	51.2% [48.1, 54.3]

The alpha test kept false promotion of edgeless trials at the rule's own rate. It cut the levered trial's rate from 35.3% to 6.3% over ten years, and it found a real edge just as often. The maintainer adopted it on 2026-11-28 as version 2 of the rule (#324). The results in this paper are for version 1.

Changing volatility and a crash. The market followed GARCH(1,1), with $\alpha = 0.08$, $\beta = 0.90$ and the same unconditional variance, and fell 20% on one random day. With no edge, the rule made 0.5% to 1.3% of trials candidates over three years, and 18.7% to 21.8% over ten. It paper-traded 0.1% to 0.3% over three years and 2.9% to 4.9% over ten. The crash lowered the candidate rate most over three years, probably because one 20% fall cuts a short record's Sharpe ratio the most. The paper-trade rate stayed below its level.

9 Discussion

The rule's two stages do different work, and in this synthetic world only the second held the false promotion of edgeless trials near its level. The DSR guards against selection among trials, yet it compares a trial with cash, so in a rising market it promotes any trial that holds the market long enough. Between 35.5% and 39% of the edgeless trials became candidates after ten years. A verdict of "candidate" therefore says that a trial's return is unlikely to be luck. It does not say that the trial has skill. The deflation by the number of trials did little, which is consistent with trials that share a benchmark and so count as fewer independent ones (López de Prado, 2019).

The second stage compares the trial with the benchmark, and it held the false promotion of edgeless trials at or below 7.5%, and at or below 9% with heavy-tailed shocks. Its test is on returns, however, and a trial that holds more of the benchmark earns more of the market's premium. Such a trial passed the rule in 29.5% of replicates over ten years, whereas the Ledoit–Wolf test of Sharpe ratios rejected it near its level. Two changes would close this gap: a difference of Sharpe ratios, or the trial's alpha after a regression on the benchmark, as the statistic inside StepM, which allows either (Romano & Wolf, 2005). Either would change the decision rule, so it is a decision for the maintainer.

The power results fit the warnings of the literature. Wiecki et al. (2016) found that backtest Sharpe ratios barely predict out-of-sample ones, and McLean and Pontiff (2016) found that published edges shrink. Here, even a rule that controls its false positives finds a modest edge rarely within the few years of data that most strategies have. Harvey and Liu (2020) argue that a test's missed discoveries matter as much as its false ones, and this calibration puts numbers on both for one rule. A project that wants to find edges with an information ratio below 1 needs longer records, more independent evidence such as a paper trade, or a rule that accepts more false discoveries.

10 Limitations

- **One synthetic world.** The market is normal with a constant premium in the first experiments. The extensions add GARCH volatility and one crash, but no regimes, no dependence between trials' active returns and no real data.
- **Correlated trials.** The trials' active returns are independent here, so the raw count of trials was the right one for the DSR. A real search reuses ideas, and its active returns correlate. This calibration doesn't measure that case. In the power experiment, the edge trial also raises the variance of the trials' Sharpe ratios, and with it the DSR's threshold.
- **Costs, but no tax or risk-free rate.** The extensions charge one cost model at one turnover. Tax needs trade-level lots, which the synthetic trials don't have.
- **Two parts are simplified.** The oracle check is assumed passed. PBO uses 8 blocks (70 splits), not the paper's 16 (12,870), to repeat it 1,000 times.
- **A small bootstrap.** StepM draws 199 samples, which makes its threshold noisier than with the 1,000 that the project's default uses.
- **Selection by Sharpe ratio only.** A researcher may select by other criteria, which this calibration does not cover.
- **Two hundred replicates in the first experiments.** Their intervals have a half-width of 2 to 7 points. The extensions' 1,000 replicates narrow that to about 1.5 to 3

points, enough to tell 6% from 10%.

11 Conclusion

In this synthetic world, the project's decision rule held the false promotion of strategies with no skill near its level. It paper-traded at most 7.5% of the best of up to 50 edgeless trials, and at most 9% when their returns had heavy tails and autocorrelation. Its first stage, the deflated Sharpe ratio, did not separate skill from market exposure, and its second stage rewarded a trial that held more of the benchmark, which a test of Sharpe ratios did not. It found a strong edge within a few years, but a modest edge rarely within ten. The rule is sound against luck in this world, and a benchmark test on risk-adjusted returns would make it sound against holding more of the market too. With costs, with a harder market and with 1,000 replicates, it still held false promotion below its level. StepM on alpha removed most of the levered trial's advantage at no cost in power.

References

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). *The probability of backtest overfitting.* Journal of Computational Finance, 20(4), 39–69. doi.org/10.21314/JCF.2016.322.

Bailey, D. H., & López de Prado, M. (2012). *The Sharpe ratio efficient frontier.* Journal of Risk, 15(2), 3–44. SSRN 1821643.

Bailey, D. H., & López de Prado, M. (2014). *The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality.* Journal of Portfolio Management, 40(5). doi.org/10.3905/jpm.[removed].

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). *Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?* Review of Financial Studies, 22(5), 1915–1953. doi.org/10.1093/rfs/hhm075.

Gençay, E. (2026). *What survives honest evaluation? Leakage-safe, search-aware assessment of LLM-driven trading strategy discovery.* Preprint. arXiv:2608.27734.

Hansen, P. R. (2005). *A test for superior predictive ability.* Journal of Business & Economic Statistics, 23(4). doi.org/10.1198/073500105000000063.

Harvey, C. R., & Liu, Y. (2020). *False (and missed) discoveries in financial economics.* Journal of Finance, 75(5). doi.org/10.1111/jofi.12951.

Ledoit, O., & Wolf, M. (2008). *Robust performance hypothesis testing with the Sharpe ratio.* Journal of Empirical Finance, 15(5), 850–859. doi.org/10.1016/j.jempfin.2008.03.002.

Lo, A. W. (2002). *The statistics of Sharpe ratios.* Financial Analysts Journal, 58(4), 36–52. doi.org/10.2469/faj.v58.n4.2453.

López de Prado, M. (2019). *A data science solution to the multiple-testing crisis in financial research.* The Journal of Financial Data Science, 1(1), 99–110. doi.org/10.3905/jfds.2019.1.099.

McLean, R. D., & Pontiff, J. (2016). *Does academic research destroy stock return predictability?* The Journal of Finance, 71(1), 5–32. doi.org/10.1111/jofi.12365.

Romano, J. P., & Wolf, M. (2005). *Stepwise multiple testing as formalized data snooping.* Econometrica, 73(4), 1237–1282. doi.org/10.1111/j.1468-0262.2005.00615.x.

Wiecki, T., Campbell, A., Lent, J., & Stauth, J. (2016). *All that glitters is not gold: Comparing backtest and out-of-sample performance on a large cohort of trading algorithms.* Working paper. SSRN 2745220.

Appendices

Appendix A: Reproduce

The simulation and the figures run from the repository's root:

```
uv run python scripts/calibrate_evaluator.py # papers/evaluator-calibration/results.jsonl
uv run python scripts/paper_figures.py # papers/evaluator-calibration/media/*.svg
```

Each replicate's seed is the text of its setting and its number, so every run gives the same results. StepM's and the Ledoit–Wolf test's own seed is the replicate's number. The rule is `open_familyoffice/verdict.py` at `PROTOCOL_VERSION` 1, which each row of `results.jsonl` records. Version 2 runs StepM on the alpha test, so the results reproduce at a commit before #324. The first experiments took about 20 minutes on 10 cores. With the extensions, the whole run took about 100 minutes on 16 workers.

Appendix B: Who did the work

An agent of the project (Claude Opus 5.5) designed the experiments, wrote the simulation and the figures, and wrote this paper. A second agent, which did not write it, reviewed its claims and citations against the review pages, and its findings are addressed in this version. Another agent wrote the extensions (#307) and their section. A second agent hasn't reviewed them yet.