

Research protocol

The standing method for a live pre-registered systematic trading research programme

This is the working protocol my own programme runs under — not a proposal or a template. 84 strategy specifications are currently recording live against it. It is included so you can see the level at which I'd expect to discuss your next run.

1. THE GOVERNING PRINCIPLE

Forward testing is free. Real money is the only hard bar. Anything plausible enters the book and records forward. No statistical test closes a line of enquiry, because a test that fails to resolve an effect has not shown the effect is absent.

Three things, and only these three, block a candidate from entering: it is **unexecutable**, it is **unmeasurable**, or it **duplicates** something already recording. Low power does not block. A weak prior does not block.

The reason is the asymmetry of the two errors. A false positive costs a slot in the book and dies on the forward record at no expense. A false negative discards a real effect permanently and silently, and nobody ever audits those. So the entry bar is low and the *trust* bar is high.

2. PRE-REGISTRATION — WHAT A SPECIFICATION MUST CONTAIN BEFORE IT RUNS

A candidate is sealed in writing before it records a single observation. The seal states:

1. **The rule**, precise enough that two people would implement it identically — including every edge case, stated explicitly rather than left to the implementation.
2. **The matched passive benchmark**. Any candidate with directional exposure must declare what it is being compared against, before results exist. Comparison against zero is not permitted, because in a rising market almost any long-biased rule beats zero.
3. **What would count as failure**, and the review point at which it will be judged — a trade count or a date, fixed in advance.
4. **The known costs of the design**, stated before the first observation. If a parameter is expected to decay, that expectation is written down so the decay is confirmation rather than discovery.
5. **The trial count** — how many variants were examined before this one was chosen.

The seal is never amended after results exist. A specification that can be clarified once it is recording is not being tested, it is being fitted. Where a genuine defect is found, the candidate is rebuilt and re-sealed as a new entry with its history disclosed — it is not edited.

3. THE RECORD

- **Append-only ledger**. Results are written once. No row is ever edited or removed. A record that can be revised is not evidence.
- **Never stop recording**. Investigation can stop at any time; recording cannot. A candidate that looks dead keeps its record, because the cost is zero and a reversal is information.
- **Protected period**. The out-of-sample window is not inspected trade-by-trade and its P&L is not analysed for concentration. The holdout's value is entirely in what it has been protected from, and every look spends some of it.

4. HANDLING MULTIPLICITY

The programme does not apply a false-discovery-rate correction. It uses a **sequential confirmatory** design instead: candidates record forward on data that did not exist when the specification was frozen, and false positives fail there without a correction being needed.

This is a deliberate trade, and the cost is stated: it buys statistical power at the price of time. A conventional correction across a book of this size would eliminate genuinely small effects along with the noise — and in these markets the real effects are small, on the order of a few basis points against a daily noise bar several times larger. What is retained from the multiple-testing discipline:

- **Trials are counted** and recorded with the result. A result quoted without its trial count is uninterpretable.
- **Plateaus are preferred to peaks.** If a parameter works at 14 but not at 13 or 15, that is noise. If a range works, that is a shape. Cheap, and it catches a great deal.
- **Tuning is permitted; the holdout is not.** Parameters may be adjusted freely on development data, and the out-of-sample retest is what adjudicates.

5. VALIDITY CHECKS APPLIED BEFORE ANY RESULT IS BELIEVED

GATE	WHAT IT CATCHES
Arithmetic geometry	Entry, exit, stop and risk relationships verified per observation. Introduced after a bug inflated a risk figure roughly twofold across a family of candidates.
Reconciliation	Trade list, equity curve and reported total must agree independently. Disagreement is treated as a defect, never reconciled by adjustment.
Too-good trip-wire	Any result above a plausibility threshold is treated as a suspected bug until proven otherwise. Favourable errors are the dangerous ones, because nothing prompts you to check them.
Lookahead	Every input confirmed available at the decision moment. A locked candidate once failed here after passing validation.
Grid arithmetic	Session and timezone offsets verified explicitly. A hard-coded minute once placed a benchmark window before the market opened and produced a fictional \$93,201 loss with no visible symptom.
Null validity	A permutation or randomisation null must centre on zero before anything derived from it is used. A circular-shift null that centred at +0.284 was discarded and replaced with per-market sign randomisation, which centred at +0.015.

6. SYMMETRIC RIGOUR — THE RULE THAT IS USUALLY MISSING

A dismissal is a claim, and carries the same burden as a discovery. Before any negative verdict is accepted, the test itself is verified: right levels, right granularity, right mechanic, output within sane bounds. False negatives are guarded as hard as false positives.

This exists because the programme caught itself dismissing effects on tests that could not have detected them. The practical form is a minimum-detectable-effect calculation: if the smallest effect a test could resolve exceeds any plausible edge, a null result from it carries no information and is recorded as *unknown* rather than *refuted*.

7. WHAT THIS PROTOCOL DOES NOT CLAIM

- It does not make a result correct. It makes a result *interpretable*, and it makes the failures visible rather than silent.
- It is slower than the alternative. Verdicts arrive in months rather than in an afternoon.
- It does not remove the need for judgement. It constrains where judgement is allowed to operate — before the run, not after it.

