RESEARCH PERSPECTIVE

From Trading Activity to Evidence of an Edge

What an execution record must show before a market-making thesis deserves confidence

Two thousand fills across 100 games can carry the statistical information of a few hundred independent observations. A live execution record shows that trading happened, and it takes much more to show that the trading earned an attractive return for its risk and capital. We set out an evidence ladder, from illustration to reconciled performance, and the tests an execution record must pass before a market-making thesis deserves confidence.

By Chris Park, Chief Investment Officer

Current edition: October 27, 2026 · 6 min read

First published October 18, 2026; revised October 27, 2026 to add key takeaways, Exhibits 1 and 2, a markout definition and the multiple-testing literature.

KEY TAKEAWAYS

01 Label evidence by level. Illustrations explain mechanisms, simulations test assumptions, live records show activity and only a reconciled record on a defined capital base shows performance (Exhibit 1).

02 Start from every decision, not every winning pair. Unfilled quotes, partial fills, cancellations and open inventory all belong in the cohort.

03 Fills cluster within games. Two thousand fills across 100 games can carry the statistical information of a few hundred independent observations (Exhibit 2).

> An evaluation framework; no fund track record or completed operating model is presented.

▼ **Method, sources and limitations**

Study type	Evidence and evaluation framework
Method	Original methodology discussion informed by primary order documentation and research on strategy selection, multiple testing and clustered data, with hypothetical arithmetic. No proprietary dataset, backtest or performance series is presented.
Data period	No empirical period or performance series is reported.
Source categories	Official execution documentation; Primary research on validation, selection and inference
Limitations	Describing a standard does not establish that all required records or controls are implemented. No audited returns, validated capacity, statistical significance or fund operating readiness is claimed.

Activity, execution quality and investment results are separate claims

A trading account can show that orders and fills occurred. Without further analysis it cannot show that the strategy earned an attractive return for the risk and capital used. A public price series is further removed still: it describes a market, not a particular account's decisions or economics. Evidence should be labeled at its actual level.

EXHIBIT 1

The evidence ladder

	What it can show	What it cannot show
Hypothetical illustration	How a mechanism works	The actual frequency, size or profitability of any outcome
Historical simulation	How a rule would have behaved under stated assumptions about fills, queues and costs	That the assumed fills were available
Live order and fill records	That orders were placed and executed, at what prices and sizes	The return on risk and capital
Reconciled performance record	Returns on a defined capital base after costs, with positions and cash tied out	Capacity at larger size, or persistence

Source: Predict Here Capital.

The levels inform one another, but they should not be blended into a single favorable narrative. In particular, an illustrative completed spread must never sit beside live-trading language in a way that invites the reader to mistake it for an actual result.

Define the intended edge as questions that can fail

Our market-making thesis divides into four hypotheses: better contract mapping may reduce false opportunities, event-aware quote selection may avoid poorly

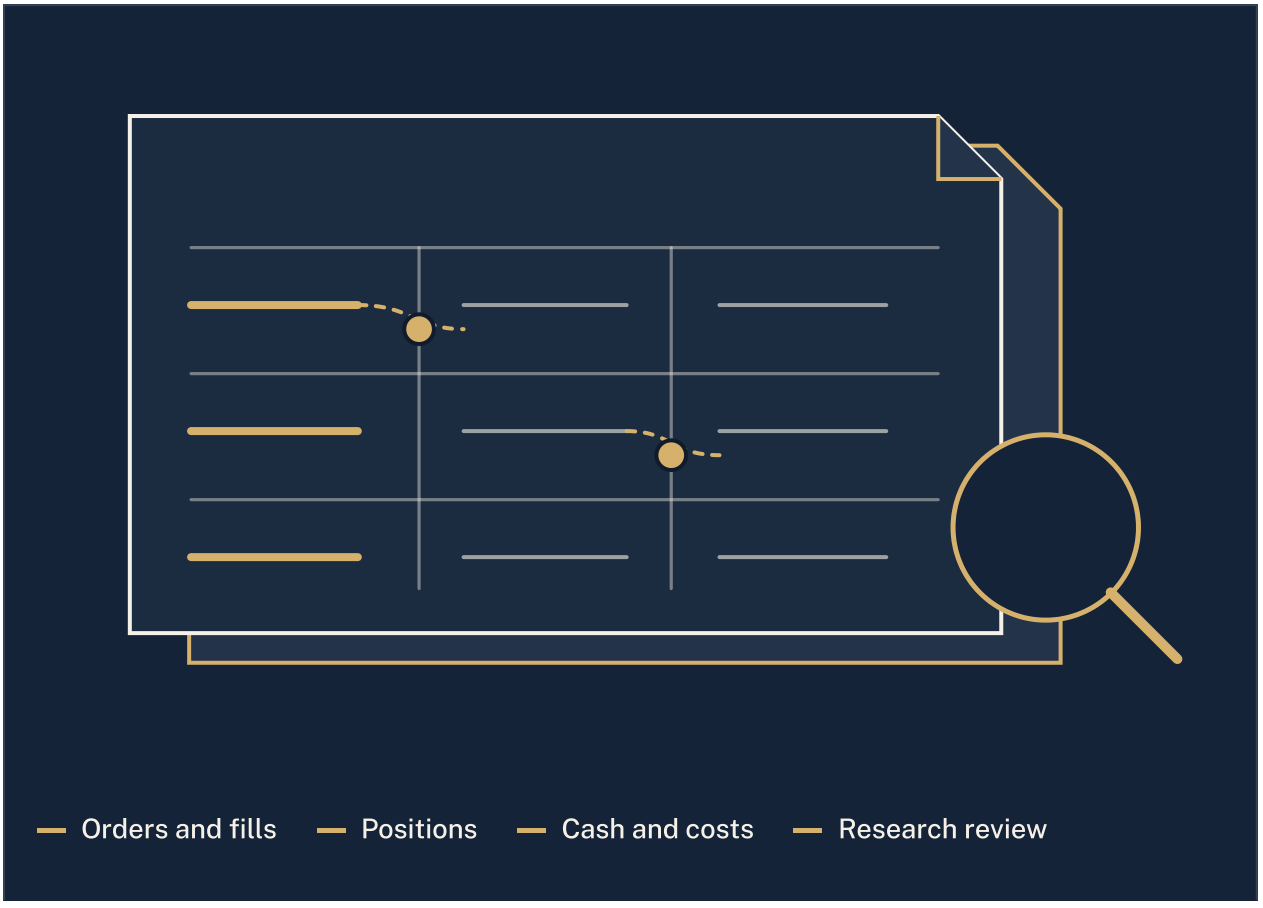
compensated exposure, inventory-sensitive decisions may reduce the cost of carrying unwanted positions, and capital allocation may improve how scarce funding is used.

Each hypothesis needs a comparison: better than which policy, over which opportunities, at what risk and capital? A lower loss rate proves little if the policy avoids almost all trading, and higher profit is not evidence of better execution if it comes from much more directional exposure.

The research should also say in advance what would weaken the explanation: an advantage that disappears after realistic costs, dependence on a temporary incentive or favorable outcomes concentrated in a handful of events.

Begin with the complete decision cohort

A study of completed winning pairs starts too late. The record begins with the universe of opportunities the policy could have considered and follows every decision about them. That includes no-trade decisions, unfilled quotes, partial fills, inventory still open at the end of the period and orders that were cancelled or rejected.



A reviewable record follows decisions through to cash and open positions.

The eligible universe and any exclusions should be defined before outcomes are known. Data outages, event cancellations and unusual conditions may justify separate analysis, but removing them because they hurt the result changes the question.

The record also needs provenance: contract terms and versions, market-data timestamps, receipt times, policy version and the information available when each decision was made. Without them, a favorable result cannot be distinguished from a hindsight reconstruction.

Use markouts as diagnostics, not as realized profits

A markout compares a fill price with a reference price a fixed interval later. It shows whether fills tend to precede adverse price movement, a common signature of being selected by better-informed counterparties.

> For a purchase filled at time t, markout at horizon h = (reference price at time $t + h$ − fill price) × quantity. For a sale the sign reverses, so a positive markout means the fill was favorable relative to the later reference price.

The reference must be stated. A midpoint, a last trade and a price available for the actual exit quantity answer different questions. A favorable midpoint markout does not mean the position could have been closed at that price, and a negative short-horizon markout does not by itself establish a final loss. Order-flow toxicity measures in the literature serve the same diagnostic purpose (Easley, López de Prado and O'Hara, 2012).

Markouts belong beside fill probability, quantity, inventory and eventual exits or settlement. They should never be added to realized P&L as if they were another source of cash, because that would double-count part of the economics.

Reconcile trades, positions and cash without hiding residual exposure

The economic record must connect orders to executions and executions to positions, and then fees, credits, settlements, corrections and transfers to account balances. A cancellation request is not proof that no further execution occurred, and an execution is not always final: Polymarket US documentation describes trades that can be voided after the fact as busted.

For an evaluation period, account value should reflect cash and remaining positions under a stated valuation policy, adjusted for external contributions and withdrawals. Realized trading gains, unrealized changes, incentives and expenses should each be identifiable, and the method must not recognize the same position twice through both trade pairing and portfolio marks.

Incentives deserve a separate reconciliation of amounts earned and amounts actually credited, with effective dates. The Polymarket US fee schedule changed on October 19, 2026, and its maker rebates now depend on the participant's assigned tier, ordinarily set by prior-month maker volume.

A credible return needs a credible denominator

A return calculated only on the capital used by completed trades misrepresents the resources needed to run the strategy. The relevant capital includes funds supporting open positions, resting orders, residual inventory and the reserves needed to keep operating through unsettled periods.

Capital use should be measured through time, not at its most favorable moment. A position that settles in minutes and one locked for several days should not receive the same capital treatment just because their final gains match.

Capacity requires another test. Larger orders may wait longer, attract different flow, consume more depth or raise the cost of reducing inventory. A small account's result

cannot establish how much capital can be deployed at comparable economics. Scaling is an execution question, not a multiplication.

Separate model selection from evaluation

A researcher who tries many quote widths, event filters, timing windows and exit rules and keeps the best-looking combination is likely to fit noise, even without deliberate cherry-picking. Bailey and co-authors show how to estimate the probability that a selected backtest is overfit, and the deflated Sharpe ratio tests whether a reported Sharpe ratio remains significant once the number of trials behind it, the length of the sample and non-normal returns are taken into account. In asset pricing, Harvey, Liu and Zhu (2016) argue that a newly discovered factor should clear a t-statistic of about 3.0 rather than the conventional 2.0, because so many have been tested. A credible report discloses the material alternatives tried and separates later evaluation from the choices that produced the policy.

Event markets add a second problem: dependence. Many fills belong to the same game or information shock, so thousands of fills are not thousands of independent observations. Exhibit 2 applies Kish's design effect for a sample mean, assuming equal-sized game clusters, a common within-game correlation and independence across games. Dividing the fill count by the design effect gives the effective sample size for that mean.

EXHIBIT 2

Clustered fills carry less information than their count suggests

Correlation of outcomes within a game	Design effect	Effective independent observations
0.05	1.95	≈1,026
0.15	3.85	≈519
0.30	6.70	≈299
0.60	12.40	≈161

2,000 fills spread evenly across 100 games (20 per game). Design effect = 1 + (20 − 1) × correlation.

Source: Predict Here Capital calculation using Kish (1965). Correlations are hypothetical.

Evaluation should reflect that dependence, report uncertainty accordingly and test whether results persist across genuinely different periods, sports and conditions.

What an investor should be able to inspect

An investor should be able to inspect the study period, universe, policy version, data provenance, costs, valuation and capital base, with figures reconciled to underlying records. Any simulated or hypothetical result should be clearly separated from actual results. Fund readiness requires separate diligence on legal documents, service providers, custody, oversight and investor eligibility. A trading result alone establishes none of those arrangements.

Sources and further reading

Sources reviewed October 27, 2026. Venue rules, fees and programs change; the specific contract and account terms govern. Views are the author's as of the edition date shown above.

1. Bailey, D. H., J. M. Borwein, M. López de Prado and Q. J. Zhu (2017). The probability of backtest overfitting. Journal of Computational Finance 20(4).
2. Bailey, D. H. and M. López de Prado (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting and non-normality. Journal of Portfolio

Management 40(5).

3. Harvey, C. R., Y. Liu and H. Zhu (2016). ...and the cross-section of expected returns. Review of Financial Studies 29(1).
4. Easley, D., M. López de Prado and M. O'Hara (2012). Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25(5).
5. Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons.
6. Polymarket US order management. Polymarket US documentation.
7. Polymarket US fee schedule. Effective October 19, 2026.