Edge Score Methodology V1

Convexly Research

2026-04-18

Edge Score: A Composite Skill Measure for Prediction-Market Traders

Convexly Research Version 1.0, 2026-04-18

Abstract

We report a composite scoring layer for prediction-market traders fit on a frozen cohort of 8,656 Polymarket wallets with at least five resolved positions. The score, Edge Score V3b, combines three standardized predictors: a posture term derived from baseline-adjusted Brier score, a conviction term derived from PnL concentration in the wallet's single largest event, and a discipline term derived from resolved position count. Under a 5-fold cross-validation with fold-local coefficient refit and fold-local standardization, the composite achieves an out-of-fold Spearman rank correlation of +0.514 with signed log PnL, against +0.147 for a Brier-only baseline. A Fama-French 2010 bootstrap null with 10,000 PnL permutations places the observed Spearman outside every permuted sample, one-sided $p < 0.0001$. Subgroup stability holds on six cross-sections. Hill alpha on realized PnL is 1.28 (95% CI 1.20 to 1.36), so the composite ranks median outcomes rather than expected returns. Two pre-registered experiments requiring per-position outcome data are deferred to V1.5.

---

1. Introduction

Calibration has a standard role in forecasting research: a calibrated forecaster, when asked to price an event at probability p, should see that event occur with frequency p. The natural question for prediction markets is whether calibrated traders earn more than uncalibrated ones. On the Polymarket profit leaderboard the answer is no. Spearman rank correlation between raw Brier score and realized PnL across 8,656 ranked wallets is +0.147. Among the top 100 wallets by profit, the relationship is stronger: worse-calibrated wallets earn more (Spearman +0.42 in a companion study). Across the full leaderboard, 85% of ranked wallets beat their own marginal-frequency baseline yet 62% of the best-calibrated quartile has negative realized PnL.

This paper defines a composite that reflects the empirical structure behind those observations. Three pillars, fit by linear regression on the full cohort, account for most of the rank-order variance in signed log PnL. The first pillar, posture, loads positively on the negation of baseline-adjusted Brier. The second, conviction, loads positively on PnL concentration in a single event. The third, discipline, loads negatively on resolved position count. The fit is reported with pre-registered validation experiments designed to answer the first question a quantitative reviewer asks of any cross-sectional score: could this be noise?

The result is not a forecasting score. A forecasting score would reward accurate probability estimation. Posture rewards the opposite direction because the training cohort's most profitable wallets are the least accurate probability estimators on their bets. The composite measures something closer to trading behaviour: which wallets in the cohort position, concentrate, and restrain position count in the pattern that correlates with profit outcomes on the shown leaderboard. The paper reports that pattern and validates it out of sample.

A related measure is Wilson (2023), which defines “Edge” as hedge fund alpha independent of VIX-short returns. The name collision is acknowledged. The construct is different: Wilson decomposes equity alpha against a volatility benchmark, while Edge Score ranks prediction-market wallets against a reference cohort percentile. The feature sets, benchmarks, and outcome variables do not overlap.

2. Data

2.1 Cohort

8,656 Polymarket wallets with five or more resolved positions as of 2026-04-15. The sample is the union of two Polymarket Data API leaderboard pulls (top-1000 and extended top-10000), deduplicated by wallet address, with a 5-position filter applied to make wallet-level Brier statistics reliable. All 9,997 wallets listed on the profit leaderboard at pull time are positive-profit by construction; the 8,656 subset is our Brier-eligible study population.

2.2 Position-level data

Per-wallet resolved positions are aggregated by (market, outcome_index) into volume-weighted average entry prices. For each wallet, we compute raw Brier score, base-rate-adjusted (skill) Brier, realized PnL, position count, and share of PnL attributable to the wallet's single largest event.

2.3 Known biases

The cohort is the Polymarket profit leaderboard. Wallets that never ranked, stopped trading, or were deactivated are not represented. The sample is a survivor cohort, cross-sectional as of a single date. Forward-looking cross-wallet validation is out of scope for V1; a within-wallet temporal holdout partially addresses this and is discussed in §5.

---

3. Method

3.1 Pillars and predictors

The composite has three pillars. Each is named for the trader archetype it captures rather than the underlying feature.

Posture is the standardized negation of baseline-adjusted Brier. Baseline-adjusted Brier is defined as observed Brier minus the wallet's own marginal-frequency Brier. Higher values mean worse calibration against baseline. Posture rewards higher values because the top-100 wallets in the training cohort are in the worst Brier quartile and account for the majority of realized profit. The pillar does not measure forecasting skill in the standard sense; it measures whether the trader makes money while calibration is imprecise.

Conviction is the standardized log of PnL concentration. Concentration is the share of total realized PnL attributable to the wallet's single largest event. Higher values denote barbell concentration: most of the wallet's return comes from one event.

Discipline is the standardized log of resolved position count, with a negative sign in the composite. Higher values of the pillar correspond to fewer resolved positions. The training cohort's most profitable wallets hold fewer, larger positions.

All three predictors are z-scored against training-slice means and standard deviations in every validation experiment (§3.3). Frozen production constants are held for the serving pipeline and are not used inside any validation fold.

3.2 Composite

Let skill denote baseline-adjusted Brier, conc denote concentration, and pos denote resolved position count. The composite raw score is:

```
raw = 0.7876 · z(−skill) + 2.7220 · z(log(conc)) − 1.1508 · z(log(pos))
```

The negation on skill is applied before standardization. The raw score is then mapped to a 0-100 percentile rank against the frozen training-cohort distribution.

3.3 Refit policy

Validation experiments refit V3b coefficients within each training fold or slice, and re-standardize predictors against training-fold means and SDs. Held-out rows are scored with the refitted coefficients and the training-fold standardization. Frozen production coefficients (0.7876 / 2.7220 / –1.1508) are used only at serving time. Per-fold coefficient drift is reported in §5 as a robustness signal.

3.4 Why V3b and not V3

An earlier variant, V3, included log of total dollars risked as a fourth predictor. V3 had a higher in-sample Spearman (+0.520) but correlated +0.30 with total dollars risked: the composite was partially measuring wallet size rather than wallet behaviour. V3b drops that term, loses roughly 10% of in-sample separation, and produces a total-dollars-risked correlation of –0.13. Capital-proxy independence is a pre-committed requirement for any shipped formula and disqualifies V3 regardless of OOS performance.

---

4. In-sample fit

Fit statistics on the full 8,656-wallet cohort, HC1-robust standard errors:

Candidate	Spearman	R^2 (HC1)	Top-bot decile PnL gap	corr(score, total_risked)
Brier baseline (V0)	+0.148	0.018	+$12,671	+0.10
V1 additive	+0.343	0.148	+$15,579	+0.13
V2 research	+0.074	0.004	+$2,759	–0.20
V3 ML (includes total_risked)	+0.520	0.270	+$15,431	+0.30
V3b ML (shipped)	**+0.420**	**0.276**	**+$10,624**	**–0.13**

Candidate	Spearman	R^2 (HC1)	Top-bot decile PnL gap	corr(score, total_risked)
V4 Kelly	+0.127	0.021	+$5,476	−0.15

Spearman values in the table are reported on the composite with the sign convention “higher composite is associated with higher PnL,” which is the reverse of the direct Spearman(raw Brier, PnL) = +0.148 cited in the companion blog post. Both describe the same underlying relationship.

---

5. Validation

Five pre-registered experiments completed on the 8,656-wallet cohort, run 2026-04-18. Two experiments requiring per-position outcome data are deferred to V1.5 and are noted where relevant. The pre-registration was published before the validation script ran.

5.1 5-fold cross-validation (E1)

Sklearn KFold with n_splits=5, shuffle=True, random_state=42. OLS fit on V3b predictors within each training fold; predictions on held-out fold accumulated into a single out-of-fold vector.

Out-of-fold Spearman with signed log PnL: **+0.514**. HC1 R^2 of OOF predictions on the outcome: 0.310. Fold-wise Spearman ranges from +0.491 to +0.535, standard deviation 0.016. Top-to-bottom decile gap in median realized PnL: +$9,055.

Per-fold coefficient stability:

Pillar	Median	Range
Posture	+0.847	[0.82, 0.93]
Conviction	+4.229	[4.14, 4.27]
Discipline	−0.772	[−0.82, −0.73]

Refit magnitudes differ from the frozen production coefficients because the standardization basis differs (fold-local vs. whole-cohort). Sign and relative ordering of the three pillars are stable across folds.

5.2 Per-wallet temporal holdout with purging and embargo (E2)

Deferred. Computing training-slice skill Brier requires per-position win/loss outcomes. The wallet_positions CSVs contain fill timestamps but not resolution outcomes, which live in a separate upstream table. A one-time join to Polymarket resolution data or a cross-venue replication on a cohort where outcomes are retrievable (e.g. Manifold) unblocks the experiment. Pre-registration stands; V1.5.

5.3 Subgroup stability (E3)

Reran the E1 protocol on six subgroups. All six clear the +0.30 pre-registered threshold:

Subgroup	n	OOF Spearman
All wallets	8,656	+0.514
Tier 2-6 (excludes top 100)	8,570	+0.511
≥20 positions	6,707	+0.562
≥50 positions	3,414	+0.589
<$10K risked	906	+0.726
≥$10K risked	7,750	+0.468

Signal strengthens as we restrict to wallets with more resolved positions, consistent with reduced noise in the underlying Brier estimates. The low-volume subgroup is small (n=906); the large Spearman on that cut should be read cautiously.

5.4 Formula-variant sensitivity (E4)

Reran the E1 protocol for all six candidates (V0 through V5) under the same fold-local refit protocol. Results:

Candidate	OOF Spearman	R^2	corr(score, total_risked)
V3 ML (all)	+0.577	0.332	**+0.192**
V1 additive	+0.521	0.308	−0.026
V3b ML (shipped)	**+0.514**	**0.310**	**−0.055**
V5 custom	+0.514	0.310	−0.055
V4 Kelly	+0.234	0.040	−0.238

Candidate	OOF Spearman	R^2	corr(score, total_risked)
V0 Brier-only	+0.147	0.018	−0.105

V3 produces the highest OOF Spearman but fails the capital- independence requirement: its correlation with total dollars risked remains +0.192 under refit, confirming that the predictor log(total_risked) operates partly as a size proxy. V3 is disqualified.

Among capital-independent candidates, V1 and V3b produce Spearman values 0.007 apart, inside V3b's own fold-to-fold noise (0.044 range). The two composites differ in a single predictor: V1 uses raw Brier, V3b uses baseline-adjusted Brier. Baseline adjustment controls for within-wallet market selection. A wallet that bets only near-certainties (resolution probabilities close to 0 or 1) records a low raw Brier score without demonstrating any forecasting skill, because the task itself is trivially easy. Subtracting the wallet's own marginal-frequency Brier removes that confound. V3b is the shipped composite. V1 is reported as the narrowly better-fitting alternative to be transparent about the sensitivity of the ranking, but the 0.007 margin is inside noise and V1 does not clear the market-selection-robustness requirement that motivated V3b's design.

5.5 Fat-tail diagnostic (E5)

Hill tail index on the positive tail of |realized PnL|. Sort absolute PnL ascending, take the largest k = 865 values (top 10%), compute alpha_hat = 1 / mean(log X_(n-i) – log X_(n-k)) for i = 0..k-1.

Alpha estimate: **1.28**. 95% bootstrap CI (500 resamples, seed 42): [1.20, 1.36]. Matches the earlier fat-tail report ($\alpha \approx 1.26$) within CI. N_eff under the Taleb approximation $N^{(\alpha/2)} \approx 331$ against a raw N of 8,656.

Alpha below 2 implies formally infinite variance on realized PnL. Edge Score rankings describe cross-sectional ordering of median outcomes. They do not bound expected returns for any individual wallet, and Kelly sizing applied to the composite under fat-tailed payoffs with uncertain edge is unsafe. See MacLean, Thorp and Ziemba (2011) on sub-Kelly behaviour and Taleb (2020) on convex-concave asymmetry under $\alpha < 2$.

5.6 Fama-French bootstrap null (E6)

The canonical skill-versus-luck test. Under the null hypothesis that the composite carries no real information, signed log PnL is permuted across wallets, V3b is refit on the permuted sample under the same fold-local protocol as E1, and the OOF Spearman is recorded. Repeat 10,000 times.

Observed OOF Spearman (E1): +0.514. Null distribution mean: ~0.0. Null 95th percentile: +0.017. Null 99th percentile: +0.025. **Zero of 10,000 permuted samples produce an OOF Spearman at or above the observed value.** One-sided $p < 0.0001$.

5.7 Information Coefficient temporal stability (E7)

Deferred. Bucketing bets by resolution quarter and computing per-period Spearman requires the same per-position outcome join as E2. V1.5.

5.8 Summary

#	Experiment	Result
E1	5-fold CV (V3b)	Spearman +0.514
E2	Temporal holdout	Deferred (data)
E3	Subgroup stability	All six subgroups ≥ +0.468
E4	V0-V5 sensitivity	V3 disqualified; V1/V3b tie (0.007)
E5	Hill α	1.28 (CI 1.20, 1.36)
E6	Bootstrap null	p < 0.0001, 10,000 permutations
E7	IC temporal	Deferred (data)

---

6. Discussion

The composite captures a cross-sectional ordering on the Polymarket profit leaderboard that survives cross-validation, subgroup cuts, and a permutation test at 10,000 iterations. The ordering is not primarily about calibration: the posture pillar enters the regression with a positive coefficient on the negation of adjusted Brier, and the best calibrated wallets on the leaderboard are not the most profitable. The ordering is primarily about conviction and discipline: wallets that concentrate PnL in a single event and make fewer, larger bets occupy the top of the realized-profit distribution on this cohort.

This is consistent with the empirical profile of the top 100 wallets studied in the companion paper. Median Brier in that group is 0.20 against a leaderboard median near 0.14. Median position count is around 20 against a leaderboard median of 50. Median concentration is 0.70. The composite encodes the shape of that profile and ranks wallets by proximity to it.

Three properties of the result matter for use. First, the bootstrap null rules out the possibility that the composite is fitting noise on a single cross-section. Second, the fold-to-fold stability of the refitted coefficients (roughly 10% range around the frozen values) argues against the fit being driven by outlier folds. Third, the fat-tail diagnostic confines all claims to ranking rather than expected-return prediction.

Two constraints bound what the composite measures. First, the training cohort is the Polymarket profit leaderboard. Wallets that never ranked are outside the training distribution; scoring them is extrapolation. Second, the fit target is realized PnL under a fixed market structure, fee schedule, and liquidity environment. If the market structure of the venue changes materially, the fit should be re-run before the scores are used operationally.

---

7. Limitations

Survivor cohort. The Polymarket profit leaderboard reports only wallets that ranked by positive profit. Wallets that traded and lost enough to drop out, or never produced enough volume to rank, are absent. This is the structural problem Taleb (2001, *Fooled by Randomness*, especially Ch. 8 on survivorship) identifies: a sample of surviving traders is not a sample of skilled traders; it is a sample of lucky-or-skilled traders with the unlucky pruned. Peters (2019) on ergodicity makes the complementary point on the cross-section versus time-average distinction: cross-sectional ensemble averages on a survivor set do not transfer to the time-average experience of an individual trader. Results here describe differences among survivors, not expected outcomes for a random new trader. V1.5 will include an active-but-unranked control cohort to measure the selection shift explicitly.

Polymarket-only fit. Cross-venue replication on Kalshi or Manifold is out of scope for V1 and is the primary target of V1.5. Methodology transfer is non-trivial: Manifold uses play money, Kalshi uses regulated dollar settlement, and Polymarket uses crypto rails. Reference cohorts will have to be built per venue.

Fat tails. Hill $\alpha = 1.28$ on realized PnL means variance is formally infinite on this distribution. Pearson correlations and OLS R^2 are not well-behaved. The paper reports Spearman rank correlation throughout as the defensible statistic. Individual realized PnL outcomes, including for high-Edge-Score wallets, remain infinite-variance.

Rank-statistic sampling distribution. A sharper version of the fat-tail critique is that any summary of OOF performance on an infinite-variance target is suspect. The distinction that makes Spearman defensible is that its sampling distribution is computed over the empirical ranks of the joint distribution, not over the raw values. Ranks are bounded by construction, so the sampling distribution of Spearman is well-behaved even when the underlying variable has infinite theoretical variance. Bootstrap confidence bands reported in §5.1 and §5.4 are on Spearman itself; at no point does the paper report a Pearson correlation, OLS R^2 confidence interval, or t-statistic on realized PnL. Parametric moment-based inference would fail under $\alpha = 1.28$; rank-based inference does not.

Permutation validity of the null model. The Fama-French bootstrap in §5.6 permutes wallet-PnL labels and recomputes the Spearman rank correlation against the held-out composite. Under the null hypothesis that Edge Score has no association with PnL, the hypothesis the test is designed to reject, wallet-PnL pairings are exchangeable by construction. The permutation distribution preserves the tail structure of the marginal PnL distribution because the operation permutes labels, not values. The tail shape is determined by the PnL marginal and is fixed across every permutation; only the wallet-to-PnL mapping varies. This is the standard Fama-French (2010) protocol applied without modification to the Edge Score context.

Selection effects in market choice. Bet selection is endogenous to the trader. A wallet that only bets near-certainties generates low raw Brier by construction. Baseline-adjusted Brier (skill Brier) partially controls for this by subtracting the wallet's own marginal-frequency Brier, but does not fully decompose skill from selection. V1 of the paper does not separate the two.

Single cross-section. The cohort is fixed at 2026-04-15. The within-wallet temporal holdout in §5.2 partially addresses temporal robustness but does not provide a forward-looking cross-wallet test. V2 of the paper plans a 30-to-90 day forward replication on a fresh cohort.

Posture nomenclature. The pillar was renamed from Calibration to Posture in V1 because the production module's user-facing pillar percentile had been flipped for intuitive readability while the composite contribution followed the fit sign. The two signals were then pointing in opposite directions. Posture aligns pillar percentile and contribution, and does not promise the pillar measures calibration precision.

V1 versus V3b margin. E4 produces an OOF Spearman margin of 0.007 favouring V1 (+0.521) over V3b (+0.514). The margin is smaller than V3b's fold-to-fold Spearman range of 0.044, meaning the candidates are not statistically distinguishable on the training cohort. V3b is the shipped composite because its calibration predictor (baseline-adjusted Brier) controls for within-wallet market selection, which the V1 predictor (raw Brier) does not. The paper reports both numbers in §5.4 to document the sensitivity; the choice is not made on the 0.007 margin.

---

8. Reproducibility

Code, raw validation outputs, anonymized cohort CSV (addresses hashed), frozen coefficients, and reference standardization constants are available on request to reviewers and collaborators. The validation runner, scoring module, and pre-registration text are all version-controlled in a private repository; a curated reproducibility bundle is distributed directly rather than hosted publicly to keep the reference cohort stable across runs.

The pre-registration text and pass-fail thresholds were committed before the validation script ran. Changes to the pre-registration (addition of E6 and E7, refit policy clarification) are recorded with timestamps and are available on request.

All random seeds are set to 42.

---

9. References

Augenblick, N., Rabin, M. (2021). *Belief Movement, Uncertainty Reduction, and Rational Updating.* Quarterly Journal of Economics 136(2), 933-992.

Fama, E. F., French, K. R. (2010). *Luck versus Skill in the Cross-Section of Mutual Fund Returns.* Journal of Finance 65(5), 1915-1947.

Forsberg, D., Gallagher, D. R., Warren, G. (2021). *Identifying Hedge Fund Skill Using Peer Cohorts.* Financial Analysts Journal 77(1).

López de Prado, M. (2018). *Advances in Financial Machine Learning.* Wiley.

MacLean, L. C., Thorp, E. O., Ziemba, W. T., eds. (2011). *The Kelly Capital Growth Investment Criterion: Theory and Practice.* World Scientific.

Peters, O. (2019). *The Ergodicity Problem in Economics.* Nature Physics 15, 1216-1221.

Taleb, N. N. (2001). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets* (2nd ed. 2004). Random House.

Taleb, N. N. (2020). *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications.* STEM Academic Press. arXiv:2001.10488.

Wilson, E. (2023). *Hedge Funds With(out) Edge: A New Measure of Hedge Fund Manager Skill.* SSRN Working Paper 4513205.