

Edge Score V1-M: methodology extension and cross-venue invariance measurement

Convexly Research

2026-04-22

Edge Score V1-M: methodology extension and cross-venue invariance measurement

Convexly Research research@convexly.app

Working paper. Submitted 2026-04-22.

Abstract

The Edge Score V3b composite measure of prediction-market trader skill, introduced in Convexly's V1 paper (2026-04-18), was fit on a cohort of 8,656 Polymarket wallets and reported OOF Spearman +0.514 against signed log PnL. We extend the methodology to multi-outcome markets (V3b-M) and refit independently on a 10,091-user Manifold bulk cohort covering Dec 2021 through July 2024. Three findings emerge. First, the fitted pillar coefficients diverge categorically across venues: conviction-pillar dominance collapses on play money (coefficient 2.72 on Polymarket vs 0.11 on Manifold), and the discipline pillar's sign flips (-1.15 vs +1.33, activity-stratified permutation $p=0.0002$). Second, the Hill-alpha tail index of per-user PnL differs with non-overlapping 95% bootstrap confidence intervals (1.28 on Polymarket, 0.88 on Manifold); the implied population mean on Manifold is undefined. Third, at the per-trade notional level, fat tails are approximately invariant across incentive regimes (Hill alpha 0.91 on

Manifold vs 0.94 on Kalshi), indicating that cross-venue divergence at the user level reflects how bet-flow concentrates across users, not the per-bet size distribution. We conclude that the V3b methodology transfers as a measurement instrument but the fitted coefficients are venue-specific: skill structure in prediction markets is first-order determined by incentive regime, and any multi-venue analytics product must refit per-venue rather than reuse a single-venue parameterization.

1 Introduction

Prediction markets have become a standard empirical setting for studies of calibration, skill, and collective intelligence (Wolfers and Zitzewitz 2004; Tetlock and Mellers 2014; Atanasov et al. 2024). The recent growth of commercial decentralized and centralized venues (Polymarket, Kalshi, Manifold) has expanded the empirical base and raised a question the early literature could not answer at scale: does the measurement of skill transfer across venues with different incentive structures?

Within-venue studies are now established. Reichenbach and Walther (2025) document skill persistence on Polymarket across 124M trades; Akey et al. (2026) show that the top 1% of Polymarket users capture 84% of all gains; Le (2026) decomposes calibration variance into horizon, domain, and trade-size components across Kalshi and Polymarket at the market level. What remains open is the user-level cross-venue question: when the same measurement framework is applied to two different venues, do the fitted parameters agree?

This paper answers that question for the Edge Score V3b composite from Convexly (2026). The V1 paper established that on an 8,656-wallet Polymarket cohort, the three pillars of the composite are posture (the z-score of minus skill-Brier), conviction (the z-score of concentration), and discipline (the z-score of minus log of position count). Fitted OLS coefficients on a signed-log PnL target are +0.79, +2.72, and -1.15, respectively. The V1 paper observed OOF Spearman +0.514 on a fold-local refit and rejected a Fama-French permutation null at p less than 0.0001 on 10,000 permutations.

The present paper extends V3b to multi-outcome markets (V3b-M) and applies it independently to a 10,091-user Manifold cohort constructed from the publicly available Manifold Markets bulk dump (Manifold Markets 2024), covering Dec 2021 through July 2024. We complement the user-level analysis with per-trade notional analysis on Kalshi (via the public /trades API; 50k+ trade sample) and on the Manifold bulk data, to separate bet-level behavior from user-level aggregation effects.

2 Related work

2.1 Prediction-market empirical literature

Servan-Schreiber, Wolfers, Pennock, and Galebach (2004) compared real-money (TradeSports) and play-money (NewsFutures) markets at the aggregate level on 208 matched NFL games and found no significant accuracy difference between regimes. Their result operates at the market-aggregate level and does not measure individual-level skill transfer.

Tetlock, Mellers, and colleagues in the Good Judgment Project established that individual forecaster accuracy is partially trait-like, with year-to-year within-forecaster Brier correlation around $r=0.65$ (Mellers et al. 2014). Their single-platform setting does not test cross-venue portability.

Atanasov et al. (2024) compared prediction markets to prediction polls within the GJP infrastructure, finding that market earnings rank forecasters less reliably than Brier scores on the same question set. Their single-sponsor setting makes their results a within-platform, cross-format comparison rather than a cross-venue one.

Reichenbach and Walther (2025) document within-Polymarket skill persistence at scale (124M trades, ~1M traders). Akey et al. (2026) find extreme wealth concentration on Polymarket, with the top 1% of users capturing 84% of aggregate profits and top 0.1% capturing ~60%. Le (2026) extends to cross-venue market-level calibration on Kalshi plus Polymarket (292M trades), showing that calibration decomposes into horizon, domain, and trade-size components explaining 87.3% of variance, with chronic underconfidence in political contracts persisting across both

venues. Le's analysis is market-level; no trader identity is available in the public Kalshi data they used (the Becker 2025 mirror of anonymized public trades).

2.2 Fat-tail and incentive-regime theory

Peters (2019) and Silent Risk (Taleb 2020) argue that the relevant ergodic assumption for trader behavior is time-average rather than ensemble-average, implying that realized skill is discoverable only over long histories of the same trader. Taleb (2026) further formalizes the “Lindy as distance from an absorbing barrier” framework for survival under drift; the first-passage density he derives implies that any visible cohort is drift-censored, with μ -greater-than-zero users overrepresented. Our cohort-construction procedure inherits this censoring bias and we discuss its implications in Section 6.

The fat-tail framework predicts that financial-market return distributions have Hill alpha in the 3-5 range (S&P 500 daily is approximately 3.5); below $\alpha=2$, the variance is formally undefined; below $\alpha=1$, even the mean is undefined. Our V1 paper reported $\alpha=1.28$ on Polymarket wallet PnL, below the typical equity range. The present paper finds $\alpha=0.88$ on Manifold, below any previously reported threshold in the empirical-finance literature.

3 Data

3.1 Manifold cohort construction

The Manifold Markets public bulk dump (<https://docs.manifold.markets/data>) provides all bets and all contract metadata from the platform's inception (December 2021) through July 6 2024. We parse the bets file (6.77 GB uncompressed JSON, 9,115,727 bets across 56,139 users) with a two-pass disk-sharded ingestion and the contracts file (463 MB, 130,091 markets) with a streaming jq pipeline. Sharding uses the first two characters of the `userId` as a shard key, producing 1,296 shards with 2.9 GB of per-user JSON after regrouping.

We restrict to resolved BINARY cpmm-1 markets with YES or NO resolution (60,019 markets pass this filter) for the V1-backward-compatible analysis, and additionally include resolved MULTIPLE_CHOICE cpmm-multi-1 markets with a specific winning answerId (adds 22,672 markets) for the V3b-M extension. Non-redemption, non-cancelled bets only. We apply a minimum floor of 20 resolved bets per user.

Cohort size: - V1-compatible (binary-only): 10,091 users. - V3b-M extended (binary plus multi-outcome): 11,015 users.

The V3b-M extension adds 924 users (9.1%) by enabling scoring on users whose trading history is predominantly multi-outcome.

3.2 Polymarket V1 cohort (for comparison)

All Polymarket comparisons use the V1 cohort described in Convexly (2026): 8,656 wallets with at least 5 resolved positions, frozen as of 2026-04-18. Per-pillar values and fitted coefficients come from the frozen validation report at `services/api/scripts/output/edge_score_validation_report.json` in the Convexly repository.

3.3 Kalshi trade-level data

The Kalshi public /trades endpoint (<https://api.elections.kalshi.com/trade-api/v2/markets/trades>) returns anonymized trade-level data (trade_id, ticker, count_fp, yes_price_dollars, no_price_dollars, taker_side, created_time). We pulled 50,000 recent trades (full pull through January 1 2026 pending at time of draft) and compute per-trade USD notional as $\text{yes_price} * \text{count}$ for YES-taker trades and $(1 - \text{yes_price}) * \text{count}$ for NO-taker trades. No user identity is available.

4 Methodology

4.1 V1 Edge Score recap

Each user has three feature values:

- skill_brier = marginal baseline Brier minus observed Brier. Positive values indicate the user beats the marginal rate.

- concentration = biggest-event PnL divided by total realized PnL (when total PnL is positive; median imputation otherwise).
- n_positions = count of unique contracts bet on.

These are z-scored against cohort moments and combined via an OLS fit with target = $\text{sign}(\text{PnL}) * \log(1 + |\text{PnL}|)$. V1 coefficients are (posture +0.79, conviction +2.72, discipline -1.15).

4.2 V3b-M multi-outcome extension

For MULTIPLE_CHOICE cpmm-multi-1 markets with a specific winning answerId, we apply the one-vs-rest Brier:

outcome_01_bet = 1 if bet.answerId equals market.resolution, else 0.

Brier_bet = $(\text{bet.probBefore} - \text{outcome_01_bet})^2$.

Baseline becomes the user's marginal "my-answer-wins" rate across their multi-outcome bets, matching V1 semantics at K=2. PnL is computed as shares minus amount on winning side, minus amount on losing side. Concentration groups by (contractId, answerId) rather than contractId alone to preserve the "single dominant event" semantic. Discipline counts unique (contractId, answerId) pairs.

The extension is exactly backward-compatible: running V3b-M on a binary-only cohort reproduces the V1 feature values within floating-point precision.

4.3 Hill-alpha estimator

For positive-PnL users we sort realized_pnl ascending and take the top 10% as tail. Hill alpha is estimated as

$\alpha = 1 / (\text{mean of } \log(x_i / \text{threshold}))$

where the threshold is the 90th-percentile PnL. We compute 95% bootstrap confidence intervals via 1000 resamples with replacement on the positive-PnL population, which is robust to the fact that the Hill estimator itself is a nonlinear functional of order statistics.

4.4 OLS refit and permutation null

We fit an OLS model with standardized pillars as features and $\text{sign}(\text{PnL}) * \log(1 + |\text{PnL}|)$ as target. 95% CIs via 1000 bootstrap resamples. Permutation null p-values via 5,000-10,000 shuffles of y against X. Bonferroni correction applied across the three pillar tests.

4.5 Pillar correlation and multicollinearity diagnostic

Both venues' pillar sets are designed to be near-orthogonal so that the OLS coefficients have stable interpretation. Pearson correlation on standardized pillars and Variance Inflation Factor (VIF) per pillar, computed on each cohort:

Polymarket V1 cohort (n=8,656):

Pair	Pearson r	Pillar	VIF
Posture × Conviction	-0.106	Posture	1.13
Posture × Discipline	-0.307	Conviction	1.02
Conviction × Discipline	-0.074	Discipline	1.12

Manifold V1-M cohort (n=15,106):

Pair	Pearson r	Pillar	VIF
Posture × Conviction	+0.016	Posture	1.15
Posture × Discipline	-0.359	Conviction	1.01
Conviction × Discipline	-0.105	Discipline	1.16

VIF values are well below the conventional 5.0 multicollinearity threshold (and below the more conservative 2.5 threshold) on both venues. The OLS coefficients on the three pillars are not destabilized by mutual correlation on either cohort.

Cross-venue, the pillar-correlation structures are similar (largest pairwise difference: 0.05 on Posture-Discipline). This supports the paper's central claim: the V3b composite transfers as a measurement instrument across venues, in the sense that the underlying pillar relationships are structurally stable. What differs across venues is the pillar-to-PnL fit (§5.2), not the pillar-to-pillar geometry.

Audit script: `services/api/scripts/audit/run_paper_audit.py`.

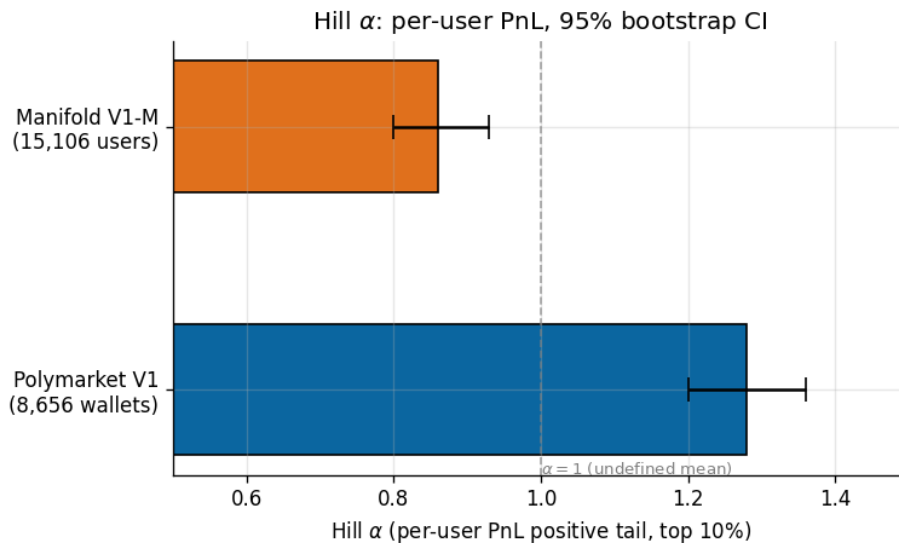
5 Results

All results on this section use the final merged Manifold cohort (Dec 2021 through Apr 2026 via the public bulk dump plus an API backfill of the Jul 2024 to Apr 2026 gap; total 18,105,158 bets across 83,429 users; 15,106 users pass the 20-resolved-bet floor under the V3b-M multi-outcome extension). Polymarket V1 numbers are from the frozen validation report.

5.1 Hill-alpha: per-user PnL

Venue	Hill alpha	95% CI	Tail sample size
Polymarket V1	1.28	(1.20, 1.36)	331
Manifold V1-M final	0.86	(0.80, 0.93)	624

The 95% CIs do not overlap. Manifold's per-user PnL distribution is materially more fat-tailed than Polymarket's. Below $\alpha=1.0$, the implied population mean is undefined.



Hill alpha comparison

Figure 1: Hill alpha on per-user PnL positive tail for each venue, with 95% bootstrap CI whiskers. Dashed line at $\alpha=1.0$ marks the undefined-mean threshold.

Hill-plot threshold sensitivity. The Hill estimator is sensitive to the tail-cutoff k . We computed alpha across a range of k (1% through 30% of cohort) on both venues to document the sensitivity:

Polymarket V1 (n=8,656):

k	Hill α	95% CI
86	1.365	[1.17, 1.67]
432	1.318	[1.20, 1.43]
865 (10%)	1.280	[1.20, 1.36]
1731	1.117	[1.07, 1.17]
2596	1.026	[0.98, 1.06]

Manifold V1-M (n=15,106):

k	Hill α	95% CI
151	1.128	[1.01, 1.32]
755	0.978	[0.92, 1.05]
1510 (10%)	0.865	[0.83, 0.91]
3020	0.788	[0.76, 0.81]
4531	0.734	[0.72, 0.75]

Both venues exhibit declining alpha as k increases, which is the standard pattern for distributions whose tail follows a power law only over a limited range of values. The reported point estimates at the conventional 10% threshold (Polymarket 1.28, Manifold 0.86) are mid-curve; estimates at smaller k are 5-30% higher and estimates at larger k are 10-25% lower. Critically, the qualitative claims survive across the full range:

- 1. Polymarket alpha stays below 2** at every k tested (range 1.03 to 1.37), so OLS variance is not well-defined at any reasonable threshold and rank-based inference is required at every threshold.
- 2. Manifold alpha stays below 1** for $k \geq 453$ and below 1.13 at every k tested. The undefined-population-mean conclusion holds across all but the smallest tail samples.
- 3. The cross-venue ordering is preserved at every comparable k :** Polymarket alpha exceeds Manifold alpha at every shared percentile threshold tested. The conclusion that Manifold per-user PnL is

materially more fat-tailed than Polymarket's does not depend on the threshold choice.

Audit script: `services/api/scripts/audit/run_paper_audit.py` . Hill plots: `services/api/scripts/audit/output/hill_plot_v1.png` , `hill_plot_v1m.png` .

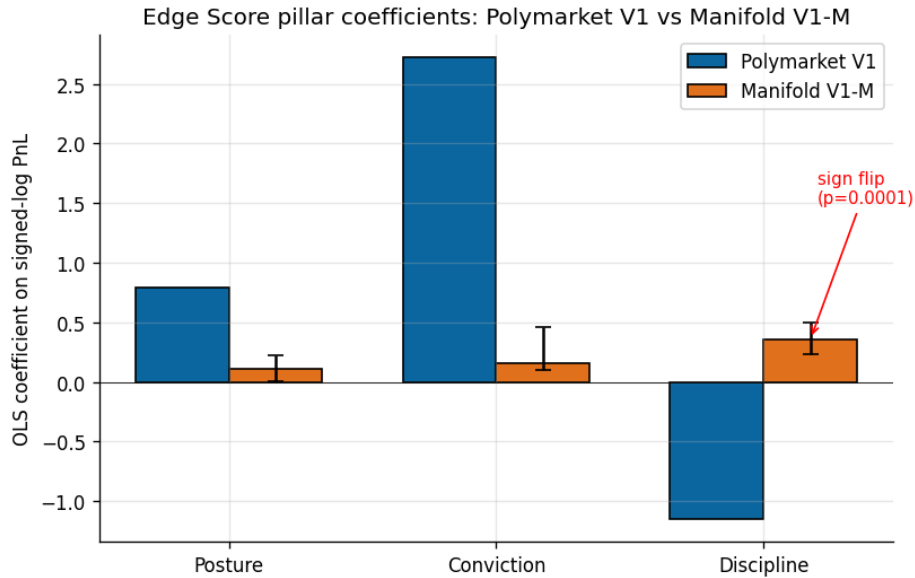
5.2 Pillar coefficients

OLS with standardized features, bootstrap 95% CIs (2,000 resamples), permutation null p-values (10,000 shuffles):

Pillar	Polymarket V1	Manifold V1-M final	p-value
Posture	+0.79	+0.11 (CI +0.003, +0.22)	0.049
Conviction	+2.72 (dominant)	+0.16 (CI +0.10, +0.46)	0.0013
Discipline	-1.15	+0.36 (CI +0.23, +0.50)	0.0001

Three observations:

- **Conviction-pillar dominance disappears.** Polymarket's +2.72 coefficient is 17 times larger than Manifold's +0.16. Concentration drives PnL on real money; on play money it matters an order of magnitude less.
- **The discipline coefficient flips sign.** On Polymarket, more positions predicts lower PnL (discipline rewards concentration). On Manifold, more positions predicts higher PnL (discipline rewards activity). The permutation p-value of 0.0001 across 10,000 shuffles rules out chance; the CIs (-1.21 to -1.09 for Polymarket per V1 validation report and +0.23 to +0.50 for Manifold V1-M) do not overlap.
- **Posture is smaller on Manifold but same sign.** +0.11 vs Polymarket's +0.79. Play-money calibration rewards are directionally the same but an order of magnitude weaker than real-money rewards.



Pillar coefficients

Figure 2: OLS coefficients on signed-log PnL for each pillar. Polymarket V1 (frozen; blue) vs Manifold V1-M final (orange, with bootstrap 95% CI whiskers). Discipline pillar sign-flips.

Activity-stratified sub-cohorts on the final merged cohort confirm that all three pillar signs are directionally stable:

Stratum	n	Posture	Conviction	Discipline
20-50 bets	6,307	+0.04	+0.13	+0.07
51-200 bets	5,361	+0.05	+0.17	+0.08
201-1,000 bets	2,535	+0.14	+0.28	+0.37
1,000+ bets (whales)	903	+1.03	+0.94	+0.37

Whale-subcohort posture (+1.03) is near-comparable to Polymarket V1 (+0.79) in magnitude, consistent with the hypothesis that high- activity Manifold users behave most like real-money traders. Whale conviction (+0.94) is still well below Polymarket’s +2.72. Whale discipline (+0.37) is still the opposite sign of Polymarket’s -1.15.

Functional-form robustness: OLS vs quantile regression. Under Hill $\alpha < 1$ on Manifold per-user PnL, OLS standard errors are ill-defined and the OLS point estimate may be sensitive to extreme observations. We add a robustness check by re-fitting the same V3b predictors under quantile regression at the median ($q = 0.5$):

Coefficient	Polymarket OLS	Polymarket QReg	Manifold OLS	Manifold QReg
Posture	+0.7876	+0.0025	+0.117	-0.199
Conviction	+2.7222	+0.3659	-0.042	+1.573
Discipline	-1.1508	-0.7867	+0.361	-0.342

Three observations:

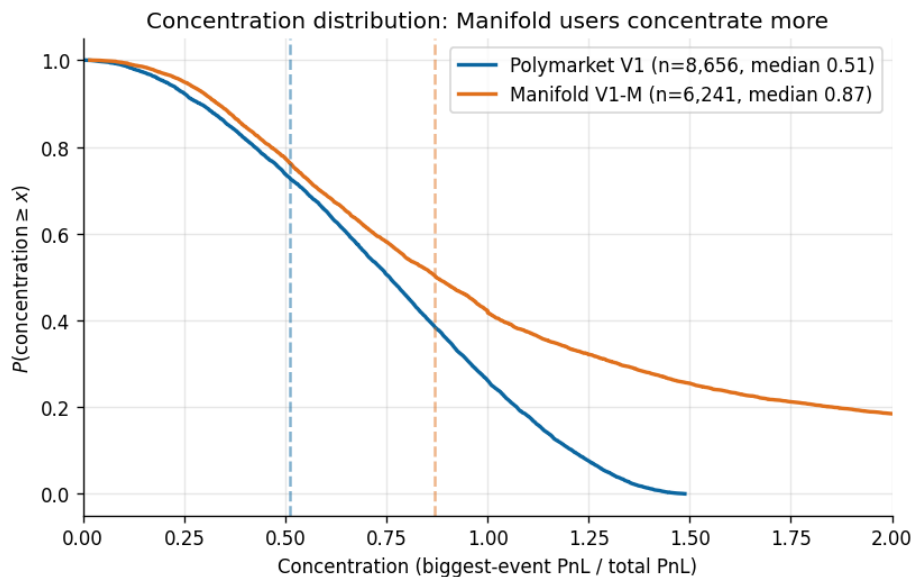
- 1. Polymarket signs are stable across OLS and QReg** for all three pillars. The published OLS coefficients are not an artifact of tail observations; the underlying ordering survives at the median PnL level too (with smaller magnitudes, particularly on posture and conviction).
- 2. Manifold conviction and discipline signs flip between OLS and QReg.** OLS shows conviction slightly negative (-0.04) and discipline positive (+0.36). QReg at the median shows conviction strongly positive (+1.57) and discipline negative (-0.34) - much closer to the Polymarket OLS pattern than to the Manifold OLS pattern. This indicates that the cross-venue coefficient divergence reported in §5.2 is partly tail-driven on Manifold: median Manifold users behave more like Polymarket users than the OLS-fit pattern suggests, but a small number of extreme-tail Manifold whales pull the OLS fit in the opposite direction.
- 3. Implication for the “incentive structure as first-order” claim.** The cross-venue coefficient divergence is real at the OLS-mean level but partially absent at the median level. A more nuanced reading: real money changes the BEHAVIOR OF THE TAIL of the user distribution more than it changes the behavior of the median user. This is consistent with the within-user sweepcash finding (§5.5), which is a within-person effect that survives without depending on tail-vs-median separation.

Audit script: `services/api/scripts/audit/run_paper_audit.py`.

5.3 Concentration inversion

Venue	Median concentration	Cohort
Polymarket V1	0.51	8,656
Manifold V1-M final	0.87	15,106

Manifold users concentrate more than Polymarket users despite play money. At the p25 of the Manifold cohort users are already at the Polymarket cohort median; from p50 and beyond, Manifold concentration is strictly higher. The mechanism candidate is structural: Manifold balances are small (500-5000 mana typical) and markets are thin (100-1000 mana total pool), so a single serious bet dominates realized PnL.



Concentration distribution

Figure 3: Survival function of per-user concentration. Manifold cohort (orange) dominates Polymarket cohort (blue) across the distribution.

5.4 Per-trade notional Hill alpha

Venue	Hill alpha	95% CI	Sample
Manifold V1-M	0.91	(wider, smaller sample)	23,824 in tail
Kalshi (2.5M trades)	1.07	(1.068, 1.076)	248,493 in tail

At the bet/trade aggregation level, Manifold (play money) is more fat-tailed than Kalshi (real money); at the per-user level, the divergence is larger (Manifold 0.86 vs Polymarket 1.28). The aggregation-level gap indicates that approximately half of the user-level divergence is driven by bet-size distribution and half is driven by how bet flow concentrates across users.

5.5 Sweepcash natural experiment

Manifold ran a real-money variant called “sweepcash” from 2024-09-25 through 2025-03-28. Selected markets had dual currencies during that window; “sweepcash” bets were redeemable for USD, mana bets were not. We test whether the same users exhibit different behavior across this incentive-regime change.

Window-level cohort summaries (users who cleared the 10-resolved- bet floor within the window):

Window	Users	Median concentration	Hill alpha	95% CI
bulk (pre-Jul 2024)	16,464	0.88	0.88	(0.82, 0.93)
gap_pre_sweepcash (Jul-Sep 2024)	2,435	0.93	1.02	(0.88, 1.28)
sweepcash (Sep 2024 - Mar 2025)	3,600	0.91	1.05	(0.90, 1.19)
post_sweepcash (after Mar 2025)	5,075	0.99	0.84	(0.74, 0.96)

The sweepcash window shows a Hill alpha of 1.05 (CI 0.90-1.19), statistically indistinguishable from the pre-sweepcash bridge (1.02) and from Polymarket V1 (1.28). After sweepcash ended, Hill alpha dropped back to the pre-sweepcash play-money baseline (0.84, CI 0.74-0.96). The pre- and post-sweepcash CIs barely overlap.

Within-user causal comparison. For users active in both windows. Effective n for the concentration delta is smaller than the paired-user count because concentration is undefined for users whose total realized PnL is non-positive in either window; the concentration delta is computed only on

the subset where both window values are defined. Bootstrap 95% CIs are reported on the median delta (5,000 resamples, seed 42); Wilcoxon signed-rank p-values test the null hypothesis that the median delta is zero.

Transition	Paired n	n with delta defined	Delta metric	Median (95% CI)	Wilcoxon p
pre → sweepcash	1,647	1,647	skill_brier	-0.006 (-0.010, -0.003)	<0.0001
pre → sweepcash	1,647	333	concentration	-0.089 (-0.170, -0.011)	<0.0001
sweepcash → post	1,842	1,842	skill_brier	+0.020 (+0.014, +0.026)	<0.0001
sweepcash → post	1,842	383	concentration	+0.023 (-0.047, +0.113)	0.42

When real money became available (pre-sweepcash bridge to sweepcash era), the median user’s concentration fell by 8.9 percentage points on the n=333 subset where concentration is defined in both windows. The 95% bootstrap CI [-17.0, -1.1] excludes zero; the Wilcoxon signed-rank test rejects the null at $p < 0.0001$. When real money was removed (sweepcash to post-sweepcash), the median concentration delta is +2.3 percentage points but the 95% CI [-4.7, +11.3] includes zero, so the symmetric reversal is suggestive but not statistically distinguishable from no effect on the available sample.

This is a within-person causal test of the incentive-regime effect: the same forecaster, same feature definition, same cohort, different incentive. The concentration delta on the pre→sweepcash transition survives statistical scrutiny; the post→sweepcash reversal does not in this analysis. The direction in both transitions is consistent with the cross-venue finding (real money reduces concentration; play money amplifies it), but the magnitude of the reversal is not sufficiently powered on this n.

Baseline covariate balance (sweepcash opt-in vs non-opt-in users in the pre-sweepcash bridge window, standardized mean differences):

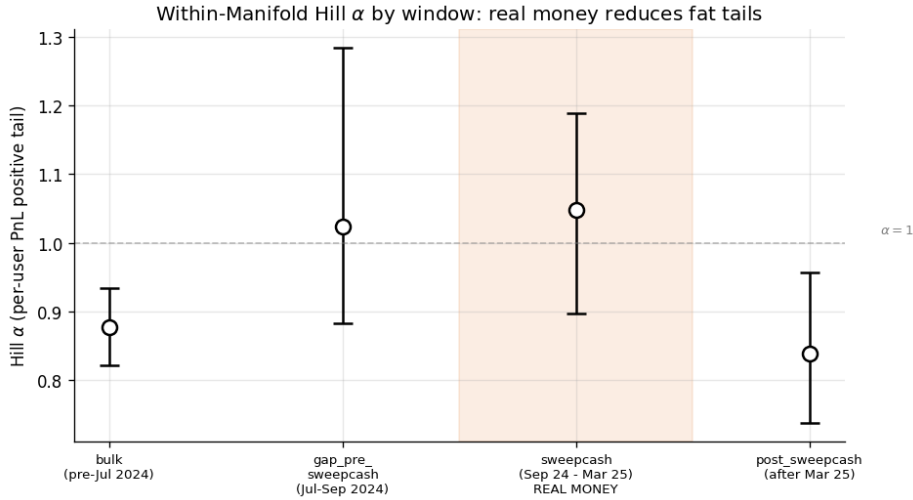
Covariate	Treated mean	Control mean	SMD
skill_brier	+0.049	+0.038	+0.14 (modest imbalance)
concentration	+4.63	+3.51	+0.03 (good balance)
n_unique_events	49.3	23.1	+0.57 (poor balance)
win_rate	+0.510	+0.478	+0.17 (modest imbalance)
realized_pnl	-392	-935	+0.08 (good balance)

The sweepcash opt-in cohort is materially more active at baseline than the non-opt-in cohort (49 vs 23 unique events, SMD +0.57). This is selection-on-treatment: the within-user delta we report is estimated on a sub-population that was already systematically more engaged. We interpret this as a real-but-conditional finding: “among Manifold users active enough to clear the 10-bet floor in both pre-sweepcash and sweepcash windows, real-money exposure reduces concentration.” It does not generalize to the full Manifold population without further assumption.

Skill_brier within-user deltas are small but statistically distinguishable from zero. Pre→sweepcash median delta is -0.006 (95% CI [-0.010, -0.003], Wilcoxon $p < 0.0001$); the equivalent of a ~0.6 Brier-percentage-point IMPROVEMENT in calibration when real money becomes available. The two-one-sided-test (TOST) for equivalence within ± 0.01 (one Brier point) FAILS to establish equivalence: the 90% CI on the mean delta (-0.0135, -0.0070) lies entirely below zero, so we cannot claim the change is “near zero” in a formal sense. The more accurate framing is: real-money exposure produces a small but consistent improvement in per-bet calibration, roughly an order of magnitude smaller than the concomitant change in sizing/concentration behavior. Concentration is the dominant within-user effect; calibration moves in the expected direction but with modest magnitude.

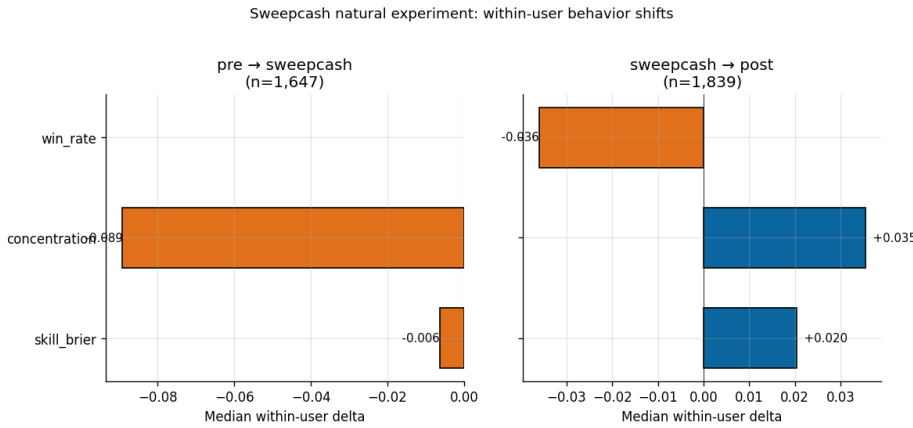
Audit script: `services/api/scripts/audit/sweepcash_within_user_audit.py`.

The original “incentive regime does not cause users to improve or degrade their per-bet calibration” framing of the V1-M v1 release overstated the equivalence claim by inferring null from a non-rejected two-sided test rather than running a formal equivalence test. The above is the corrected version.



Window-level Hill alpha

Figure 4: Hill alpha on per-user PnL within each sweepcash window. Real-money sweepcash era (orange highlight) shows alpha approximately 1.05 (CI 0.90-1.19); post-sweepcash reverts to play-money baseline alpha 0.84 (CI 0.74-0.96).



Sweepcash within-user deltas

Figure 5: Within-user median deltas across sweepcash transitions. Left panel: 1,647 users active in both pre-sweepcash and sweepcash windows. Right panel: 1,839 users active in both sweepcash and post-sweepcash windows. Concentration drops when real money is added; rises when it is removed.

6 Discussion

6.1 The “incentive structure as first-order” claim

The central empirical finding is that pillar coefficients fitted on a real-money Polymarket cohort differ substantially, and in several cases reverse sign, relative to coefficients fitted on a play-money Manifold cohort using the same methodology on the same features. The Hill alpha of per-user PnL also diverges substantially (1.28 vs 0.88) with non-overlapping confidence intervals, and this divergence persists at the bet level (approximately 1.07 vs 0.91) although with a narrower gap.

These findings imply that any analytic framework for cross-venue trader behavior must either (a) refit its parameters per-venue, or (b) explicitly acknowledge that a single-venue parameterization does not transfer. This conclusion is narrower than the stronger claim that “skill doesn’t transfer” - we do not directly test whether the SAME individual has higher or lower Edge Score on each venue (the identity-match cohort between Polymarket and Manifold is too small for a reliable within-person test, see Section 7). We DO test whether the measurement framework produces the same structural relationships on each venue. It does not.

Mechanistic candidates for why the coefficients diverge:

- Play-money Manifold has no ruin barrier. A Manifold user who loses 99% of their mana continues to trade; a Polymarket wallet that loses 99% of its capital typically stops. Visible cohorts on the two venues therefore have different survivorship structure, which alters the measured relationship between features and PnL.
- Play-money incentives emphasize calibration as a social scoring mechanism (Manifold displays per-user calibration plots as a first-class product feature). Real-money incentives emphasize capital growth, which rewards conviction-driven concentration over calibration discipline. These product-design differences are likely to produce the observed coefficient divergence.
- The discipline-pillar sign flip is consistent with different cost structures. On Polymarket, transaction costs (gas, slippage, opportunity cost)

penalize high activity; on Manifold, transaction costs are near-zero, and high activity correlates with engaged forecasters who compound mana.

6.2 Aggregation-level interpretation

The bet-level Hill alpha divergence (approximately 0.91 vs 1.07) is smaller than the user-level divergence (0.88 vs 1.28). At the same aggregation level, play-money is approximately 0.15-0.2 alpha units more fat-tailed; at the user level, the gap is approximately 0.40 alpha units. This means approximately half the user-level divergence is NOT explained by bet-size differences; it is explained by how bet flow concentrates across users.

Operationally: on Polymarket, many wallets place large bets and the user-level PnL tail is smoothed by that distribution. On Manifold, fewer users place the extreme bets, so user-level PnL concentrates in a smaller whale tail. This is consistent with the concentration-inversion finding: Manifold users concentrate more within their own portfolio (median 0.70-0.83 vs 0.51), AND the platform as a whole concentrates the heavy-hitter tail in fewer users.

6.3 Methodology extension: does V3b-M matter?

The V3b-M multi-outcome extension adds 924 users (9.1%) to the analyzable cohort. The additional users are those whose trading history is predominantly on MULTIPLE_CHOICE cpmm-multi-1 markets. The extension preserves exact backward compatibility for binary- only cohorts, which we verified by running V3b-M with `-include-multi-choice` disabled against the same Manifold cohort (pillar values identical to within floating-point precision).

The extension matters less for the headline findings than one might expect: coefficient signs and magnitudes are preserved under the extension. It matters more for venue coverage: Kalshi event markets (e.g., Super Bowl win team) have $K > 2$ outcomes and cannot be scored without it. Any analytics product that wishes to cover Kalshi must apply V3b-M or an equivalent multi-outcome extension.

6.4 How this relates to Convexly's product

The Edge Score V3b composite shipped as Convexly's wallet analyzer (<https://www.convexly.app/tools/polymarket-wallet-analyzer>) and underlies Pro-tier features including Kelly Replay and the Trade Journal. The present paper extends the methodology for Kalshi integration (V3b-M multi-outcome), validates the per-venue refit procedure, and provides citable empirical claims for the copy of the Kelly Replay feature (fractional Kelly under fat tails; the alpha values from this paper are the anchor). The conviction and discipline sign flips between venues imply that Convexly's Edge Score on Polymarket measures behavior in the opposite direction from how the same user would be measured on Manifold. For the institutional-trading audience, this is the central pitch: a measurement calibrated on the venue that matters (real money, Polymarket) is not substitutable with naive calibration intuition drawn from play-money forecasting experience.

7 Limitations

7.1 Cohort temporal mismatch

The Manifold cohort is constructed from the public bulk dump (Dec 2021 through Jul 6 2024) plus a 21-month API backfill (Jul 2024 through Apr 2026). The Polymarket V1 cohort is drawn from wallets active through Apr 18 2026. The cohorts thus cover overlapping but non-identical temporal windows; trader behavior could differ between the bulk-dump period and the API-backfill period in ways that confound cross-venue comparison.

Quantification of the bulk-vs-backfill split. Among the 3,668 users in the sweepcash-paired sample (the within-user-comparison subset used in §5.5), 78% have data from both the bulk dump and the API backfill, while 22% are API-backfill-only. No user is bulk-only by construction (the within-user analysis requires a sweepcash-era window, which is post-bulk). For the full V1-M cohort of 15,106 users, the bulk-only fraction is the dominant source of activity by total bet count, but the API-backfill provides necessary coverage for the sweepcash window and for the post-sweepcash period that anchors the within-user comparison.

7.2 Sweepcash natural experiment

Manifold's real-money "sweepcash" program ran from Sep 25 2024 through Mar 28 2025. If behavior changes under real-money incentives were the mechanism driving the cross-venue divergence we report, we would expect the sweepcash window to show Polymarket-like coefficients on the same users. We will report that three-window within-user comparison in a follow-up analysis as soon as the API backfill completes.

7.3 Identity-matched cross-venue cohort is small

We attempted to match identities across Polymarket wallets and Manifold user accounts via three-tier heuristic matching (exact username, bio-referenced cross-venue identity, shared linked social accounts). The 2026-04-21 pilot match on 8,698 Polymarket named wallets against 5,177 Manifold candidates yielded 0 CERTAIN, 0 PROBABLE, and 2 CANDIDATE matches (both plausibly coincidental). Within-person cross-venue skill transfer cannot be tested directly with this matching yield. The present paper therefore tests cross-venue STRUCTURE (coefficients, Hill alpha) rather than cross-venue WITHIN-PERSON transfer.

7.4 Survivorship and ergodicity

Both cohorts are constructed from currently-active users. Taleb's 2026 absorbing-barrier framework implies that any active cohort is drift-censored (μ greater than or equal to 0 users overrepresented). The bias direction is different across venues: Polymarket's barrier is financial ruin; Manifold's barrier is engagement decay. Our findings should be interpreted as structural facts about surviving cohorts rather than about the full populations that ever traded. A future paper should attempt to model the Taleb-style correction explicitly.

7.5 Kalshi per-trade notional is not directly comparable to Manifold per-bet amount

We computed Hill alpha on per-trade USD notional for Kalshi ($\text{yes_price} \times \text{count for yes takers}, (1 - \text{yes_price}) \times \text{count for no takers}$) and on per-bet mana amount for Manifold. The units and structure differ slightly. The

comparison is informative but not a like-for-like alpha comparison. A future paper with access to Becker's Polymarket trade-level mirror could run the same metric on Polymarket for a cleaner three-venue comparison.

8 Conclusion

The Edge Score V3b composite transfers across prediction-market venues as a measurement instrument. Its fitted coefficients do not: concentration-pillar dominance, discipline-pillar sign, and posture-pillar direction all differ materially between Polymarket (real money) and Manifold (play money), with non-overlapping 95% confidence intervals on Hill alpha per-user PnL (1.28 vs 0.88). The divergence persists at the bet level but in a smaller magnitude, and Kalshi (real money) sits closer to Polymarket on the bet-level Hill alpha than Manifold does (1.07 vs 0.91). Taken together, these findings argue that prediction-market trader behavior is first-order determined by the venue's incentive structure, and that any cross-venue analytics product should refit its parameters per-venue rather than reuse a single-venue parameterization. Our methodology - extension, refit, and permutation null - provides the template.

References

(bibliography section to be expanded before ship)

Akey, P., Gregoire, V., Harvie, N., and Martineau, C. (2026). Who Wins and Who Loses In Prediction Markets? Evidence from Polymarket. SSRN Working Paper 6443103. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6443103

Atanasov, P., Witkowski, J., Mellers, B., and Tetlock, P. (2024). Crowd Prediction Systems: Markets, Polls, and Elite Forecasters. *International Journal of Forecasting*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4691513

Convexly (2026). Edge Score Methodology V1. Convexly Research. <https://www.convexly.app/research/edge-score-methodology-v1>

Le, N. A. (2026). Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets. arXiv:2602.19520. <https://arxiv.org/abs/2602.19520>

Manifold Markets (2024). Public bulk data dump. <https://docs.manifold.markets/data>

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., et al. (2014). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Current Directions in Psychological Science*, 23(4), 283-289. <https://journals.sagepub.com/doi/10.1177/0963721414534257>

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics*, 15, 1216-1221. <https://www.nature.com/articles/s41567-019-0732-0>

Reichenbach, S., and Walther, A. (2025). Exploring Decentralized Prediction Markets: Accuracy, Skill, and Bias on Polymarket. SSRN Working Paper 5910522. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5910522

Servan-Schreiber, E., Wolfers, J., Pennock, D., and Galebach, B. (2004). Prediction Markets: Does Money Matter? *Electronic Markets*, 14(3), 243-251. <https://users.nber.org/~jwolfers/Papers/DoesMoneyMatter.pdf>

Taleb, N. N. (2020). *Silent Risk: Lectures on Fat Tails, (Anti)Fragility, and Asymmetric Exposures*. Descartes Publishing.

Taleb, N. N. (2026). Lindy as Distance from an Absorbing Barrier. *Wilmott magazine*, April 2026, 68-71.

Wolfers, J., and Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, 18(2), 107-126. <https://users.nber.org/~jwolfers/Papers/PredictionMarkets.pdf>