

## **7531498-large-orders-liquidity-and-price-adjustment-in-prediction-markets**

**SUMMARY:** The article investigates how large orders affect the liquidity and price adjustment in prediction markets. It uses high-frequency Polymarket trades and order books to study the distinction between price adjustment and incremental information. The study finds that short-run quote changes increase with order size relative to displayed depth, but this relation is weakened when episode isolation is strict. Prices are closer to eventual outcomes five minutes after large-order episodes, but this descriptive improvement is not larger than after comparable ordinary trades. The article also suggests that exact historical books show that book-implied slippage ranks realized execution costs, and subsequent reversal is concentrated in selected event-proximate sports markets. The findings indicate that large orders can cause larger price changes without significantly improving forecast loss. The study distinguishes liquidity-sensitive quote adjustment from incremental outcome information in prediction markets.

**METHOD:** The research uses high-frequency Polymarket trades and order books to analyze the relationship between large orders, displayed liquidity, and price adjustment. A 28-day prospective sample of Polymarket trades is used, with exact historical books for analysis. The study isolates episodes of large orders and compares their impact on price adjustment with ordinary trades.

### **KEY FINDINGS:**

- Short-run quote changes increase with order size relative to displayed depth.
- Prices are closer to eventual outcomes five minutes after large-order episodes.
- Exact historical books show that book-implied slippage ranks realized execution costs.
- Subsequent reversal is concentrated in selected event-proximate sports markets.

**IMPLICATION FOR TRADING:** The findings suggest that large orders can cause larger price changes without significantly improving forecast loss. This distinction between liquidity-sensitive quote adjustment and incremental outcome information is crucial for understanding how prediction markets function. Practitioners can use these insights to better understand market dynamics and improve their trading strategies.

# Large Orders, Liquidity, and Price Adjustment in Prediction Markets

Constantin J. Schneider\*

September 27, 2026

## Abstract

Prediction-market prices aggregate information through trading, but aggressive orders also consume finite displayed liquidity. I study the distinction between price adjustment and incremental information using high-frequency Polymarket trades and order books. In a 28-day prospective sample, short-run quote changes increase with order size relative to displayed depth. Strict episode isolation materially weakens this relation. Prices are closer to eventual outcomes five minutes after large-order episodes, but this descriptive improvement is not larger than after comparable ordinary trades. A model that adds signed order flow to the prevailing price and unsigned depth pressure likewise produces no detectable out-of-time improvement in forecast loss. Exact historical books show that book-implied slippage ranks realized execution costs, while subsequent reversal is concentrated in selected event-proximate sports markets. The evidence separates liquidity-sensitive quote movement from incremental outcome information.

**Keywords:** prediction markets, market microstructure, liquidity, forecast performance, order flow, price discovery.

**JEL Classification:** G12, G14, D82.

---

\*University of Münster, Germany. [constantin.schneider@wiwi.uni-muenster.de](mailto:constantin.schneider@wiwi.uni-muenster.de).

# 1 Introduction

Prediction markets convert heterogeneous beliefs into tradable prices. A binary contract that pays one dollar if an event occurs and zero otherwise trades between zero and one, so a price of 0.60 is often read as a 60 percent probability. This interpretation makes prediction markets useful in politics, macroeconomic forecasting, public health, corporate planning, and geopolitical risk. A sharp move following one large order may be read as new information even when the order mainly trades through a shallow book. This ambiguity raises a central empirical question: when prices move over short horizons, do they reflect information aggregation or order-flow pressure?

Prediction markets combine features of bounded-payoff derivatives, limit-order markets, and political-risk instruments. Their prices enter media coverage, portfolio decisions, and policy discussions even when the underlying contracts are thin. Market quality therefore depends on how ex-ante prices perform against eventual outcomes and on how short-run quotes respond to concentrated trading in finite-depth books. This distinction connects prediction markets directly to the finance literature on liquidity, price discovery, and market efficiency.

I ask two connected questions. First, how strongly do prediction-market prices respond when large trades absorb a substantial share of displayed liquidity? Second, does the direction of those trades improve forecasts beyond the information already contained in prices and unsigned order size? The distinction matters because the same price change can reflect information, execution against a thin book, or both. I therefore study the event-time relation among concentrated order flow, displayed depth, and price movement. I then compare forecast loss after large and ordinary trades and test whether signed order flow adds out-of-time predictive content. I do not observe the latent event probability, so the outcome tests concern realized forecast performance and not probability truth at the trade timestamp.

The paper’s main contribution is to distinguish liquidity-sensitive quote adjustment from incremental outcome information in prediction markets. Combining prospective depth measures, a market-disjoint outcome sample, and exact book replay, the evidence shows that large orders are followed by larger price changes when displayed depth is low, without a detectable improvement in outcome forecasts relative to comparable ordinary trades. This separation clarifies why short-run quote movement need not signal superior information.

I combine transactions with prospective and historical orderbook states, settlement outcomes, matched timestamps, pre-event news counts, complete on-chain fill histories for resolved markets, and exact Kalshi–Polymarket contract pairs. The first prospective collection measures displayed depth before live Polymarket episodes over 24 hours. A broader 28-day collection yields 173,240 valid five-minute episodes from 2,271 markets. Joint trade-and-orderbook coverage spans 20.71 days in 19 segments. A later market-disjoint sample covers August 14 through September 1. It supplies 19,072 isolated episodes for the pre/post comparison and 20,167 resolved episodes for the

out-of-time prediction exercise, both from 1,706 markets. The third-party PMXT archive<sup>1</sup> permits exact pre-trade replay for 36 event hours from April through July 2026.

The 28-day prospective sample provides the broadest direct depth test. Moving depth pressure from 0.25 to 2 predicts 1.610 percentage points of five-minute directional movement, or 1,409 dollars per 100,000 dollars of exposure, with a restricted wild-cluster  $p$ -value of 0.0001. Strict same-token isolation reduces the estimate to 0.659 percentage points, or 577 dollars, with a 95 percent bootstrap- $t$  interval from 298 to 854 dollars. The depth-pressure relation remains positive at 15 minutes but is not statistically detectable after one hour. The sign is stable across alternative thresholds, outlier treatments, and recorder-coverage checks, although recorder gaps affect magnitude and selective missingness cannot be ruled out.

Prices move closer to outcomes after large orders, but a matched ordinary-trade benchmark shows a nearly identical improvement, and I do not detect an out-of-time forecast improvement from signed order flow.

Exact book replay shows that book-implied slippage ranks realized execution costs and that subsequent reversal is concentrated in event-proximate sports markets.

Wallet-level settlement outcomes do not support broad persistent trader alpha, although estimates vary across cohorts.

Cross-platform tests show limited same-direction price co-movement, but stricter controls substantially attenuate the relation. Common information remains an alternative explanation.

Together, the evidence supports a limited interpretation. Large-order episodes on Polymarket are followed by liquidity-sensitive price movements, while the outcome tests do not detect incremental forecast information specific to those orders. The calibration change is not statistically detectable, and the design does not identify a causal liquidity effect. Exact historical reversal is concentrated in selected sports settings. Wallet and cross-venue evidence does not establish persistent trader skill or an independent order-size transmission effect. Prediction-market prices aggregate information through a trading mechanism, and their interpretation should account for both outcome performance and the liquidity that supports the quoted probability.

Prediction markets use prices to aggregate dispersed information into event-contingent forecasts. Wolfers and Zitzewitz (2004) survey this forecasting role, while Berg and Rietz (2006) and Berg, Nelson, and Rietz (2008) document evidence from the Iowa Electronic Markets. Betting markets also predated modern polling in the United States and internationally (Rhode and Strumpf, 2004, 2008). Historical U.S. presidential markets often produced informative election forecasts (Rhode and Strumpf, 2004). Forsythe et al. (1992) show that political stock markets can aggregate information despite biases among individual traders. Forsythe et al. (1999) review broader evidence that election markets can forecast efficiently even when individual traders are biased and error-prone.

A prediction-market price is an equilibrium object, so its short-run movement need not be a pure update in beliefs. Manski (2006) shows that market prices need not equal mean beliefs in

---

<sup>1</sup>The archive is available at <https://archive.pmxt.dev/Polymarket/v2> (accessed September 10, 2026). PMXT distributes the archive under CC BY 4.0.

a simple equilibrium model. Gjerstad (2005) and Wolfers and Zitzewitz (2006) characterize how the mapping from beliefs to prices depends on risk preferences, the distribution of beliefs, and trader wealth. Page and Clemen (2013) find that prices are reasonably calibrated near expiration but exhibit a favorite-longshot pattern at longer horizons. A quoted probability therefore embeds beliefs, risk preferences, budget constraints, transaction costs, liquidity, and market design.

A broad empirical literature studies forecasting performance. Servan-Schreiber et al. (2004) find that play-money and real-money markets were similarly accurate in their NFL sample. Using a broader comparison, Rosenbloom and Notz (2006) find that the real-money market was more accurate for nonsports events and examine differences related to trading volume and event type. Spann and Skiera (2003) evaluate virtual stock markets in three business-forecasting applications. In a laboratory estimation task, Graefe and Armstrong (2011) find no overall significant accuracy difference between prediction markets and several other group methods, with prediction markets underperforming face-to-face groups on some questions. Polgreen et al. (2007) show that a market populated by health-care workers predicted statewide seasonal influenza activity two to four weeks in advance. Taken together, these studies document forecasting performance that varies across incentive schemes, tasks, and information environments.

Political forecasting work is especially relevant because event-market prices are often used as public signals. Erikson and Wlezien (2008) challenge claims that election markets dominate polls and show that properly discounted poll-based forecasts outperform both vote-share and winner-take-all market prices in their Iowa Electronic Markets sample. Erikson and Wlezien (2012) extend the comparison to historical U.S. elections and find that, once polls are available, market prices add no predictive information beyond polls and largely follow them. Prediction-market price changes can therefore combine information, liquidity, and strategic order flow. They need not represent clean belief revisions.

A second literature studies manipulation and strategic trading. Hanson (2003) develops market mechanisms for aggregating probability assessments. In a controlled experiment, Hanson, Oprea, and Porter (2006) find that manipulation incentives did not reduce price accuracy because other participants adjusted their trading behavior. This evidence is reassuring when markets are liquid, competitive, and populated by traders willing to trade against distorted prices.

The reassuring result is less complete in thin markets. A price can be locally manipulable even when it remains informative in expectation. Work on outcome manipulation makes this point directly. Chakraborty and Das (2014) show that prediction markets change incentives when participants can influence the underlying outcome. Recent theory on leveraged event contracts further separates market-price manipulation from outcome manipulation (Nechepurenko, 2026b). I study market-price pressure, specifically the association between concentrated order flow and changes in the public probability signal. I do not observe intent or identify outcome manipulation.

In Kyle (1985), competitive market makers infer information from aggregate order flow. Glosten and Milgrom (1985) show how adverse selection shapes bid-ask spreads and transaction prices. Hasbrouck (1991) documents that a trade's price impact can arrive with a lag and varies with

trade size. Limit-order-market evidence links liquidity, spreads, depth, and quote revisions after large trades (Biais, Hillion, and Spatt, 1995). Prediction markets differ from equity markets because the asset payoff is event-contingent and bounded between zero and one. The economic logic still applies. If a large order is substantial relative to available depth, the trade can move the price even without a corresponding change in fundamental probability.

Market-maker design reinforces this logic. The logarithmic market scoring rule acts as a continuously available automated market maker with bounded subsidy costs, and its scale parameter governs how sharply prices respond to orders (Hanson, 2007). This matters for the interpretation of modern event markets. If a large trader can move the public signal at low cost, the price is no longer only a probability estimate. It is also an object that strategic actors may try to influence, especially when prices are cited by journalists, campaigns, or financial commentators.

Recent work on Polymarket and Kalshi extends this literature to modern event-contract platforms. Le (2026) documents domain- and horizon-specific calibration patterns and finds that the large-trade calibration pattern in political contracts observed on Kalshi is not robust on Polymarket. I examine a complementary object: within-episode changes in realized forecast loss around large orders, together with contemporaneous order size relative to displayed depth. Qin and Yang (2026) introduce an on-chain trade archive and show how transaction-classification errors propagate into microstructure and forecast-performance measures. Nechepurenko (2026a) studies fill-side concentration in a short legacy-market window and emphasizes the attribution limits of public execution records. Dai, Jia, and Yu (2026) show that short-horizon settlement design can create manipulation incentives and settlement-time distortions in the underlying market. Shen et al. (2026) document a consistency gap between off-chain matches and reverted on-chain settlements. My design connects these modern institutional concerns to the classical microstructure distinction between information-based price discovery and liquidity-sensitive price adjustment.

Temporary price pressure predicts a larger immediate response when order size is high relative to displayed depth, followed by limited persistence or reversal. Informed trading instead predicts persistent directional changes and lower realized forecast loss. A stable identity-based information advantage could also appear in lagged wallet performance. These contrasting predictions organize the empirical tests.

## 2 Prediction-Market Structure and Data

### 2.1 Platforms and Data Sources

Polymarket and Kalshi allow users to trade event-contingent claims. Many contracts are binary. A Yes position pays one dollar if the event occurs and zero otherwise. A No position pays one dollar if the event does not occur and zero otherwise. Market prices are therefore commonly interpreted as implied probabilities, subject to fees, spreads, collateral constraints, and market microstructure. Polymarket records settlement and ownership through blockchain-linked infrastructure. Kalshi

operates a separate centralized event-contract order book. The platforms do not share trader identifiers or a matching engine.

The platform is useful for studying large-trader effects because it exposes several complementary layers of data. Market records describe contracts, trading conditions, and resolutions. Transaction records provide trader proxies, direction, size, price, and timing, while blockchain identifiers permit consistent matching across sources. Position data identify large current holdings. Current order-book snapshots report best bids, best asks, spreads, and depth. Price histories provide high-frequency contract-level observations.

Kalshi’s public historical interfaces report market definitions, settlement outcomes, transaction timestamps, contract counts, prices, and taker direction. The cross-platform design matches these contracts to Polymarket using titles, named outcomes, event dates, and contract language without using price behavior. It then links every matched Kalshi contract to the corresponding Polymarket Yes-token price history. The separate venues make the comparison informative about transmission beyond one matching engine, although public news can affect both venues simultaneously.

## 2.2 Sample Construction

Table 1 separates the seven empirical samples. Panel A is a retrospective convenience sample of 100 active markets collected from paginated public endpoints. The trade request retrieves at most 200 recent trades per market. The resulting 19,993 trades therefore do not constitute a complete history or a volume-weighted platform sample. Panel B is a separate prospective collection that records live trades and order books for 50 selected active markets over 24 hours. Panel C is the 28-day prospective sample. Its rolling universe contains up to 50 active markets at each refresh, and 2,271 distinct markets contribute valid five-minute outcomes. Joint recorder coverage equals 20.71 days across 19 segments, and the largest gap lasts 2.06 days. Panel D is a later outcome sample covering August 14, 2026, 12:06:49 UTC through September 1, 2026, 23:59:59 UTC. It excludes every condition observed in the earlier samples and is chronologically later and market-disjoint. The pre/post comparison requires a five-minute quote and retains 19,072 resolved episodes. The forecast comparison targets settlement and does not require that later quote, so it retains 20,167 resolved episodes. Because the split was selected retrospectively, I do not describe it as a preregistered holdout. Panel E contains complete on-chain fill histories for 50 resolved standard binary markets. The 20 highest-volume eligible markets were analyzed first. The 30-market out-of-sample group was selected from lower volume ranks before its outcomes were collected. No profitability measure entered the selection rule.

Panel F contains 321 resolved Kalshi contracts with exact Polymarket counterparts. The collection begins from Kalshi politics and election series ranked by historical volume and retains pairs only when event, outcome, and timing agree. A manual audit found no semantic mismatch among the 92 pairs added in the final expansion. Eighty-one of these pairs have a decisive terminal Polymarket price at or below 0.01 or at or above 0.99. All 81 agree with Kalshi’s settlement. The

Table 1: Sample construction

| Object                                                  | Count       |
|---------------------------------------------------------|-------------|
| <i>Panel A: Retrospective active-market sample</i>      |             |
| Active Polymarket markets                               | 100         |
| Trades                                                  | 19,993      |
| Holder observations                                     | 9,385       |
| Current endpoint orderbooks                             | 200         |
| Token-level price observations                          | 288,191     |
| Whale-candidate trades                                  | 1,964       |
| Whale episodes                                          | 1,511       |
| Episodes with matched 60-second prices                  | 440         |
| Episodes with matched one-hour prices                   | 386         |
| <i>Panel B: Prospective 24-hour orderbook sample</i>    |             |
| Recorded markets                                        | 50          |
| Trades inside recording window                          | 37,454      |
| Orderbook snapshot rows                                 | 235,400     |
| Whale-candidate trades                                  | 13,953      |
| Whale episodes                                          | 4,124       |
| Episodes matched to pre-trade depth                     | 4,065       |
| Markets in matched event-time sample                    | 48          |
| <i>Panel C: Prospective 28-day sample</i>               |             |
| Raw recorded trades                                     | 5,533,328   |
| Unique valid trades                                     | 4,627,776   |
| Whale episodes                                          | 352,086     |
| Episodes matched to pre-trade depth                     | 191,249     |
| Episodes with valid five-minute outcomes                | 173,240     |
| Markets with valid five-minute outcomes                 | 2,271       |
| Joint recorder coverage in days                         | 20.71       |
| <i>Panel D: Later, market-disjoint outcome sample</i>   |             |
| Pre/post episodes before resolution filtering           | 20,011      |
| Resolved pre/post episodes                              | 19,072      |
| Forecast episodes before resolution filtering           | 21,355      |
| Resolved forecast episodes                              | 20,167      |
| Resolved markets                                        | 1,706       |
| Event families                                          | 1,284       |
| <i>Panel E: Resolved-market profitability sample</i>    |             |
| Resolved standard binary markets                        | 50          |
| Complete on-chain fill events                           | 3,769,374   |
| Transactions                                            | 1,508,984   |
| Unique wallets                                          | 356,315     |
| Pre-specified out-of-sample markets                     | 30          |
| Out-of-sample fill events                               | 1,238,910   |
| <i>Panel F: Cross-platform resolved-contract sample</i> |             |
| Exact Kalshi–Polymarket contract pairs                  | 321         |
| Kalshi trades                                           | 739,814     |
| Binary payout exposure in dollars                       | 265,420,300 |
| Polymarket five-minute prices                           | 6,140,564   |
| Conservative q95 macro-event clusters                   | 33          |
| <i>Panel G: Historical exact-orderbook validation</i>   |             |
| Selected event hours                                    | 36          |
| Downloaded context hours                                | 144         |
| Filtered archive events                                 | 5,671,153   |
| Reconstructed pre-trade books                           | 1,384       |
| Exact event-study episodes                              | 1,317       |
| Quote-eligible episodes                                 | 1,055       |

remaining eleven histories stop before a decisive terminal quote and are retained because their contract definitions are exact matches.

Panel G uses the PMXT Polymarket order-book archive. It combines 32 event hours selected without reference to price outcomes with four high-notional stress hours. For each event hour, the sample includes the preceding hour and two following hours. The archive is a third-party capture of Polymarket’s public order-book feed and is not an official historical endpoint.

## 2.3 Measurement and Coverage

The unit of price analysis is the traded token. This matters because a market can contain both Yes and No tokens. A purchase of Yes and a sale of No are both directionally positive for the event. To avoid mixing mechanically opposite exposures, all price changes are converted into directional changes. For a Buy episode, a later price increase is positive. For a Sell episode, a later price decrease is positive. Directional changes are reported in price units. A change of 0.001 equals 0.1 percentage points in implied probability.

Historical depth is unavailable from the platform endpoint used for the retrospective sample. The first prospective collection addresses this limitation for one 24-hour window by storing repeated orderbook snapshots before live episodes. In that matched sample, the median pre-trade snapshot age is 18.35 seconds and the maximum is 52.45 seconds. The 28-day sample applies the same strict pre-trade timing rule in a rolling top-50-volume universe. It contains 5,533,328 recorded trade rows and 4,627,776 unique valid trades, which form 352,086 large-order episodes and 173,240 five-minute regression episodes. Because recorder restarts create 19 joint segments and several multiday gaps, the sample expands prospective market and calendar support without observing a continuous 28-day platform panel. The PMXT data reconstruct the last strict pre-trade book from archived full snapshots and ordered price changes. This design improves timestamp precision and supports a direct movement-and-reversal decomposition, but it remains selected by event hour and market activity and does not represent the full platform.

I use external sources as measurement checks alongside the prospective sample. In the locally overlapping hour, PMXT and the recorder agree on trade direction for all 13,513 common transaction and asset combinations, match trade size within tolerance in every case, and reproduce all 22 depth measures for 702 exactly aligned order-book states. These conditional matches support measurement consistency but do not establish that the PMXT archive is complete. The recorder and event-driven archive use different sampling schemes, and only 0.561 percent of recorder book states share an identical state identifier. PMXT does not provide wallet information.

I separately validate the direction and wallet attached to trades from Polymarket’s data interface against the exchange events recorded on Polygon. The check covers 1,000 transactions from the later sample. All 829 token purchases and 171 token sales match the corresponding on-chain taker-order event in transaction hash, wallet, token, direction, and quantity. Appendix Table A1 reports the reconciliation. This check supports the measurement of signed order flow. It does not establish that the trades are informed or that the recorded wallets identify economically independent traders.

News measurement has two parts. The first is a descriptive audit that combines Google News RSS windows with manual relevance and direction coding for 100 episodes selected after observing event-study magnitudes. The second queries Google News for the 33 historical validation market-hours and counts only articles published during the exact 24 hours before the event. The pre-event count enters the historical regressions directly. Both measures remain title-based and cannot capture every public signal. Live scores and scheduled event progression are particularly important omissions in sports markets.

Price-history data are also incomplete for some older episodes in the active-market sample. The main control and pre-trend tests therefore use strict matched-price windows. Episodes without complete price windows for a given test are excluded from that test.

### 3 Large-Trade Measures and Empirical Strategy

The empirical strategy has three layers. The primary price-response design links large-order episodes to displayed depth before the first trade and measures directional midpoint changes after the final trade. The same specification is estimated in an initial 24-hour sample and in a 28-day prospective sample, with strict same-token isolation and multiple post-event horizons. A later, market-disjoint sample evaluates prices against eventual outcomes. I first report the pre/post Brier comparison, then benchmark it against comparable ordinary trades and test whether signed order flow adds out-of-time forecast information. Exact historical PMXT books evaluate book-implied slippage against realized execution costs and separate immediate movement from subsequent reversal. Matched timestamps, pre-trends, news, settlement performance, prior wallet histories, and cross-venue price changes serve as supporting tests of alternative interpretations.

A trade enters the baseline candidate set if its notional is at or above the 95th percentile of trades in the same market or if its wallet is at or above the 95th percentile of cumulative wallet notional in that market. The wallet measure is computed over the complete collection window and is therefore a retrospective activity classification, not an ex ante trader label. The union rule is intentionally broad. In the 24-hour prospective sample, 2,606 of 4,124 episodes qualify through the wallet criterion without containing a 95th-percentile trade. The 28-day sample applies the same within-sample q95 union rule and reports large-trade-only, q90, and q99 specifications. Neither definition identifies manipulative intent or private information.

Candidate trades are clustered into episodes. An episode groups trades with the same market, wallet, side, outcome, and token when consecutive candidate trades occur no more than 300 seconds apart. Because the rule applies to consecutive trades, total episode duration can exceed 300 seconds. The retrospective sample contains 1,511 episodes from 1,964 candidate trades. The 24-hour prospective sample contains 4,124 episodes from 13,953 candidate trades. In the 28-day sample, 1,639,324 candidate trades form 352,086 episodes. Of these, 191,249 have a valid pre-trade snapshot no more than 60 seconds old and 173,240 have a valid five-minute outcome.

The later outcome sample uses the same q95 large-order rule, 300-second episode construction, quote tolerances, and strict same-token isolation. It spans August 14, 2026, 12:06:49 UTC through September 1, 2026, 23:59:59 UTC and excludes every market observed in the earlier samples. The temporal split was designated retrospectively, so I call it the *later, market-disjoint outcome sample* and do not describe it as a preregistered holdout. Markets tied to the same underlying event are grouped into event families using Polymarket’s event identifier. Inference clusters at that level. Observations with missing or conflicting family identifiers are excluded and remain visible in the attrition counts.

For resolved binary market  $m$ , episode  $i$ , and time  $s \in \{pre, post\}$ , let  $p_{ims}$  denote the traded token’s midpoint and  $y_m \in \{0, 1\}$  its realized payoff. Episode-level Brier loss is  $BS_{ims} = (p_{ims} - y_m)^2$  (Brier, 1950). The outcome estimand averages  $BS_{im,post} - BS_{im,pre}$  within each market and then equally across markets, so active and inactive contracts receive the same total weight. The post price is measured 300 seconds after the isolated episode. A negative estimate means lower average realized squared forecast error. Inference uses event-family clustering and 9,999 restricted wild-bootstrap draws.

The accompanying calibration diagnostic assigns prices to ten fixed-width bins using  $b = \min\{\lfloor 10p \rfloor, 9\}$ . Each episode receives the reciprocal of its market’s retained episode count. Within each bin, miscalibration equals the bin’s share of total weight multiplied by the squared difference between the weighted mean forecast and weighted outcome. The reported measure sums this quantity across bins (Murphy, 1973). I report this calibration diagnostic separately from average Brier loss because the two measures capture different aspects of forecast performance.

The pre/post comparison alone cannot attribute an improvement in forecast loss to large orders. Information may arrive during any five-minute interval, particularly near the resolution of sports contracts. I therefore construct a risk set of ordinary trades that are observed in the same token and are comparable in starting probability, time to settlement, recent market activity, volatility, spread, and event time. Every included large-order episode has at least one ordinary-trade comparison with a valid five-minute quote and no overlapping large-order window. I estimate propensity scores in this common-support sample and use overlap weights. The estimand is the market-equal difference between the five-minute Brier change after large orders and the corresponding change after ordinary trades. Standard errors are clustered by event family, and the two-sided restricted wild test uses 9,999 event-family sign flips. This comparison is not causal because unobserved information arrival may still differ across the two groups.

I next test whether the direction of a large order improves forecasts beyond information already embedded in the market. The baseline logistic model uses the pre-episode price and unsigned depth pressure. The augmented model adds signed depth pressure, which is positive for token purchases and negative for token sales. Both models are estimated only in the earlier 28-day sample. Their coefficients and transformations are then held fixed and applied to the later market-disjoint sample. I compare market-equal Brier loss from the two sets of predictions and cluster inference by event family. The temporal separation prevents refitting on the evaluation markets. Because the exercise

was developed after outcomes for the later period had become available, I do not present it as a preregistered forecast competition.

The signed-flow analysis evaluates all 20,167 resolved episodes with the required forecast inputs. Settlement prediction does not require an additional five-minute quote. As a robustness check, I impose the five-minute-quote requirement used in the pre/post comparison, which yields 19,072 episodes. The conclusion remains unchanged. The risk-set analysis requires every large-order episode to have at least one eligible ordinary-trade comparison and measures recent activity using all valid prior trades. I also consider a broader specification that retains additional large-order episodes and constructs activity only from the control-candidate sample. It yields an estimated Brier-loss difference of -0.0017 with a 95 percent confidence interval from -0.0038 to 0.0005. Across these alternative specifications, I do not detect a large-order-specific improvement in forecast accuracy.<sup>2</sup>

For the historical event study, the event timestamp is the beginning of the episode. Post-event horizons are 60 seconds, five minutes, 15 minutes, one hour, six hours, and 24 hours. Pre-event horizons are 60 seconds, five minutes, 15 minutes, and one hour. These estimates describe movement after the first episode trade and may include continued trading within multi-trade episodes. For the prospective depth design, the pre-event midpoint and displayed depth are measured before the first trade, while post-event horizons begin at the final trade. Total episode notional can therefore be scaled by pre-episode depth without using future trades to explain an earlier price change. The outcome variable is the directional change in the traded token. A buy is positive when the token price rises, and a sell is positive when the token price falls.

The matched-timestamp design compares whale episodes to ordinary timestamps in the same token. Controls are selected away from whale windows and matched on the nearest observed base price. In the baseline specification, each treated episode receives five controls, and timestamps within one hour of any whale episode in the same token are excluded. The main sample contains 382 treated episodes and 1,910 assigned control rows with complete price windows through one hour. These rows correspond to 614 unique token-time pairs, and some control timestamps are assigned to more than one treated episode. The comparison holds fixed the token and starting price but not the exact information environment, spread, or intraday volatility.

The matched comparison is therefore a descriptive benchmark and not a causal counterfactual. The news audit provides a second descriptive check by separating no-news windows, noisy-news windows, ambiguous relevant news, news in the whale direction, and news against the whale direction among selected large movements. Neither design rules out public or private information arriving at the episode timestamp.

The matched-timestamp bootstrap resamples treated episodes after pairing each episode with the mean of its assigned controls. This procedure gives each treated episode equal weight but does not remove dependence created by reused control timestamps. The resulting intervals are reported as a descriptive sensitivity analysis. The matching grid varies the number of controls per episode

---

<sup>2</sup>The final sample and support definitions were selected after preliminary analyses. Under the five-minute-quote restriction, the signed-flow estimate is 0.0001 with a 95 percent confidence interval from -0.0003 to 0.0005. I report the alternative constructions to make the sensitivity to these choices transparent.

across 1, 3, 5, and 10 and the exclusion window across 30 minutes, one hour, two hours, and six hours.

The pre-trend test compares movement before and after the whale episode. If whale episodes merely follow an existing trend, prices should move in the whale direction before the episode. If the movement is generated by price pressure at the event, the post-event movement should exceed the pre-event movement. This test does not prove causality, but it directly addresses the immediate momentum concern.

The profitability analysis separates quote movement from trader gains. Midpoint profit is measured from the volume-weighted episode price to the future midpoint. Best-quote profit closes a Buy episode at the future best bid and a Sell episode at the future best ask. This measure is optimistic when displayed quantity at the best quote is smaller than the episode. A full-size specification therefore retains only snapshots whose relevant top-of-book depth covers the entire episode size. All profit measures are gross of platform and gas fees and are reported in dollars per 100,000 dollars of binary payout exposure.

The ordinary-trade benchmark applies the same future-quote calculation to all 37,454 trades in the prospective recording window. Regressions compare whale and ordinary trades with token and UTC-hour fixed effects, controls for trade direction, entry price, and squared entry price, and standard errors clustered by market. The preferred large-trade comparison does not control for trade size because the treatment itself is defined by the market-specific size distribution. None of the 146 markets successfully matched from the active-market samples was resolved on July 13, 2026, so this component remains a short-horizon marking exercise.

The separate resolved-market analysis uses complete on-chain fill histories. For each fill, realized profit equals the collateral flow plus the final token payout, net of recorded exchange fees and before gas costs. Payout exposure is the number of binary shares. An observation is classified as a whale order under the preferred definition when either its fill notional or its wallet’s cumulative notional in that market is at or above the within-market 95th percentile. This classification is retrospective and describes trade size or wallet activity. It is not an ex ante measure of information. The analysis reports exposure-weighted profit levels and market-equal whale-minus-nonwhale differences. Student- $t$  intervals treat markets as independent observations. Role assignment identifies one aggressive taker event per transaction and all opposing events as maker fills. Maker and taker payout exposure balances exactly within every market.

Three ex-ante designs remove future activity from the large-order classification. The initial-period design uses the first 20 percent of distinct trading timestamps in each market to estimate an order-size threshold and a cumulative wallet-notional threshold. These thresholds are fixed before evaluating the remaining timestamps. The expanding design divides distinct trading timestamps into ten ordered blocks. The first block is excluded from evaluation. For every later block, thresholds and prior wallet activity are recalculated using only earlier blocks. An order satisfies the union definition when its current size exceeds the historical size threshold or its wallet’s previously ob-

served market notional exceeds the historical wallet threshold. Separate size-only and wallet-only specifications identify which component generates the result.

The Kalshi extension uses an initial-period trade-size design because the venues do not share wallet identities. Within each matched Kalshi contract, the first 20 percent of distinct trade timestamps determine q90, q95, and q99 contract-count thresholds. These initial trades are excluded from every outcome. Later trades at or above the threshold form the large-trade group. Other later trades form the benchmark. Settlement P&L is calculated from the transacted contract price and final binary payout. Differences are exposure weighted within each contract and averaged equally across event clusters.

Short-horizon outcomes preserve the Kalshi taker’s directional exposure. The local outcome compares the Kalshi transaction price with the first transaction at or after five minutes, 30 minutes, or six hours. A price is retained only when it occurs within ten minutes of the target horizon. The cross-platform outcome uses the latest Polymarket Yes price within ten minutes before the Kalshi trade and the first Polymarket price within ten minutes after the target horizon. Every effect is the large-minus-other difference in marked value per 100,000 dollars of binary exposure. Same-day congressional primaries within California, Georgia, and New York are collapsed into state-level macro-events. Inference uses 9,999 event-level Rademacher sign flips. This design identifies lead-lag predictability, not a causal effect of Kalshi order flow, because common public information can move both venues.

The third design tests persistent trader identity across markets. At the start of each target market, it identifies wallets above the 95th percentile of cumulative notional in non-overlapping market histories that ended before the target market began. Target-market fills are then classified using this fixed prior-market wallet label. The design requires at least five earlier market histories. It uses no target-market activity to define the wallet group.

The role-specific interaction is the market-equal difference between the whale taker effect and the whale maker effect,

$$\hat{\beta}_{WT} = \frac{1}{M} \sum_{m=1}^M [(\hat{r}_{W,T,m} - \hat{r}_{N,T,m}) - (\hat{r}_{W,M,m} - \hat{r}_{N,M,m})].$$

This quantity is the interaction coefficient from a saturated whale-by-role comparison with market effects and equal market weights. Student- $t$  intervals use the cross-market distribution. Restricted wild inference applies 9,999 market-level Rademacher sign flips.

## 4 Empirical Results

The historical event study provides descriptive evidence of short-run movement after episode initiation. Of 1,511 episodes, 440 have matched prices at 60 seconds, five minutes, and 15 minutes. The available sample falls to 386 episodes after one hour and 255 after 24 hours. Mean directional movement equals 0.080 percentage points after 60 seconds, 0.098 percentage points after five min-

Table 2: Episode-level event study

| Horizon    | Episodes | Mean change | Probability points |
|------------|----------|-------------|--------------------|
| 60 seconds | 440      | 0.0008      | 0.080              |
| 5 minutes  | 440      | 0.0010      | 0.098              |
| 15 minutes | 440      | 0.0010      | 0.104              |
| 1 hour     | 386      | 0.0011      | 0.107              |
| 6 hours    | 369      | 0.0005      | 0.046              |
| 24 hours   | 255      | -0.0026     | -0.256             |

Table 3: Prospective event-time orderbook depth

| Depth pressure | Episodes | Pressure | 60s pp | 5m pp | 1h pp  |
|----------------|----------|----------|--------|-------|--------|
| $\leq 0.25$    | 3,657    | 0.022    | 0.008  | 0.008 | -0.011 |
| 0.25 to 1      | 271      | 0.519    | 0.081  | 0.009 | -0.098 |
| 1 to 2         | 75       | 1.370    | 0.308  | 0.241 | 0.149  |
| $> 2$          | 62       | 6.243    | 0.894  | 0.864 | 1.118  |

utes, 0.104 percentage points after 15 minutes, and 0.107 percentage points after one hour. The 24-hour mean is negative, but the changing sample prevents interpreting this estimate as clean within-episode reversal.

The average magnitudes are modest but economically interpretable. In a binary prediction-market token, 0.001 equals 0.1 percentage points in implied probability. A one-hour effect of 0.107 percentage points changes the mark-to-market value of a 100,000 dollar position by about 107 dollars. In liquid markets, this is small relative to total volume. In thin markets, the same movement can be large relative to spread, displayed depth, and the public signal produced by the price.

Cross-sectional splits in the retrospective sample provide suggestive evidence on market frictions. The highest-spread quartile has a one-hour directional movement of 0.431 percentage points, while the lowest-liquidity quartile has a one-hour movement of 0.372 percentage points. High-liquidity markets show little short-run movement. These unadjusted comparisons motivate a direct depth test but do not by themselves distinguish spread, volatility, information, and available orderbook depth.

The strongest orderbook evidence comes from the prospective live collection. The recorder stores orderbook snapshots for 50 active markets and 102 monitored tokens, together with live trades, from July 11, 2026 at 05:21 UTC to July 12, 2026 at 05:22 UTC. This produces 235,400 snapshot rows and 37,454 trades inside the recording window. Applying the candidate and clustering rules yields 4,124 episodes. Of these, 4,065 episodes from 48 markets can be matched to a pre-trade orderbook snapshot within 60 seconds.

Table 3 reports movement from the pre-episode midpoint to horizons measured after the final episode trade. Depth pressure is total episode notional divided by top-5 notional depth on the relevant pre-trade book side. Episodes below one quarter of top-5 depth move by an average of 0.008 percentage points after five minutes. Episodes above twice top-5 depth move by 0.864

Table 4: Prospective depth-pressure regressions after episode completion

| Horizon    | Baseline |         | Market FE |         | Wallet cluster |         | Large trade |         | N     | Dollars |
|------------|----------|---------|-----------|---------|----------------|---------|-------------|---------|-------|---------|
| 60 seconds | 0.449*** | (0.096) | 0.453***  | (0.065) | 0.453***       | (0.108) | 0.392***    | (0.076) | 4,056 | 397     |
| 5 minutes  | 0.396*** | (0.115) | 0.410***  | (0.046) | 0.410***       | (0.147) | 0.386***    | (0.068) | 4,042 | 359     |
| 15 minutes | 0.330**  | (0.129) | 0.384***  | (0.127) | 0.384***       | (0.147) | 0.288**     | (0.112) | 3,995 | 336     |
| 1 hour     | 0.588*** | (0.174) | 0.674***  | (0.099) | 0.674***       | (0.164) | 0.587***    | (0.143) | 3,824 | 590     |

Notes: Each cell reports the coefficient on log one plus top-5 depth pressure, with its standard error in parentheses. The baseline uses HC1 standard errors. Market FE adds market and UTC-hour fixed effects and clusters by market. Wallet cluster uses the same fixed effects and clusters by wallet. Large trade restricts the sample to episodes containing a 95th-percentile trade and clusters by market. Dollars report the preferred implied value change per 100,000 dollars of exposure when depth pressure rises from 0.25 to 2. \*\*\*, \*\*, and \* denote significance at the 1, 5, and 10 percent levels.

percentage points. The mean difference is 0.856 percentage points, or 856 dollars per 100,000 dollars of exposure. The high-pressure mean is based on 61 episodes with five-minute outcomes. Its median is 0.200 percentage points, equivalent to 200 dollars per 100,000 dollars of exposure. These episodes occur in six markets, and three markets account for 79 percent of the group. The bin comparison is therefore economically large but should not be read as the effect for a typical episode.

Table 4 reports regression estimates. The baseline model controls for log episode notional and the pre-trade spread. The preferred specification adds market and UTC-hour fixed effects and clusters standard errors by market. The five-minute depth-pressure coefficient is 0.410 percentage points. Clustering by wallet instead yields a standard error of 0.147 and a two-sided  $p$ -value of 0.005. Restricting the sample to episodes that contain a 95th-percentile trade produces a coefficient of 0.386 percentage points. Moving from depth pressure of 0.25 to 2 changes log one plus pressure by 0.876. The preferred estimate therefore implies a 0.359 percentage point movement, or 359 dollars per 100,000 dollars of exposure. This regression-adjusted magnitude is less sensitive to the small and concentrated high-pressure bin than the raw mean difference. Appendix Tables A3 and A5 assess sensitivity to episode composition, outliers, market concentration, candidate thresholds, episode windows, and orderbook freshness.

Table A4 addresses event overlap and finite-cluster inference. An episode is isolated only if no other same-token episode has a trading interval that intersects the window from the focal episode's first trade through the relevant post-event horizon. At five minutes, this restriction retains 956 of 4,042 observations across all 48 markets. The coefficient falls from 0.410 to 0.244 percentage points, with a market-clustered standard error of 0.138 and a restricted wild-cluster  $p$ -value of 0.042. The standardized movement from depth pressure 0.25 to 2 is 0.214 percentage points, or 214 dollars per 100,000 dollars of exposure. Only two isolated five-minute episodes exceed twice displayed depth. The standardized value should therefore be interpreted as a regression contrast, not as a high-pressure group mean. Restricted wild-cluster  $p$ -values for the full sample are 0.001 at 60 seconds, below 0.001 after five minutes, 0.131 after 15 minutes, and 0.052 after one hour. The evidence is strongest at the immediate horizons and does not establish persistent impact.

Table 5: Depth pressure and directional price movement in the 28-day prospective sample

| Horizon    | Sample   | Coef. pp | Cluster SE | Wild $p$ | Episodes | Markets |
|------------|----------|----------|------------|----------|----------|---------|
| 60 seconds | Full     | 2.214    | 0.151      | 0.0001   | 185,235  | 2,342   |
| 60 seconds | Isolated | 0.483    | 0.096      | 0.0001   | 56,251   | 2,158   |
| 5 minutes  | Full     | 1.610    | 0.151      | 0.0001   | 173,240  | 2,271   |
| 5 minutes  | Isolated | 0.659    | 0.163      | 0.0001   | 32,976   | 1,804   |
| 15 minutes | Full     | 1.516    | 0.207      | 0.0001   | 152,676  | 2,065   |
| 15 minutes | Isolated | 0.548    | 0.177      | 0.0011   | 16,124   | 1,149   |
| 1 hour     | Full     | 1.662    | 0.315      | 0.0001   | 100,717  | 1,388   |
| 1 hour     | Isolated | 0.464    | 0.338      | 0.1476   | 5,003    | 619     |

Notes: Coef. pp is the coefficient on log one plus top-5 depth pressure in percentage points. Models control for the pre-trade spread, log episode notional, market fixed effects, and UTC-hour fixed effects. Episode notional is floored at one dollar before taking logs. Cluster SE is clustered by market. Wild  $p$  reports the two-sided null-restricted Rademacher wild-cluster bootstrap- $t$   $p$ -value using 9,999 replications, where 0.0001 is the smallest attainable corrected value. An episode is isolated when no other same-token episode intersects the window from its first trade through the stated post-event horizon. Post-event horizons begin after the final episode trade.

#### 4.1 Price Response in the 28-Day Prospective Sample

The 28-day sample tests whether the 24-hour relation appears in a strictly later and substantially broader prospective collection. It uses the q95 union candidate rule, a 300-second episode window, the latest pre-trade snapshot within 60 seconds, and a five-minute post-episode outcome with a 60-second matching tolerance. The regressions measure top-5 directional depth and include market and UTC-hour fixed effects, the pre-trade spread, and episode notional. Inference clusters by market and uses 9,999 null-restricted Rademacher wild-cluster bootstrap replications.

Isolation separates a focal episode from nearby trading in the same token. An episode is classified as isolated only when no other same-token episode overlaps the interval from the focal episode’s first trade through the relevant outcome horizon. Because this restriction changes both sample size and support, the difference between the full and isolated estimates is an attenuation comparison, not a causal decomposition.

Table 5 reports the 28-day estimates. At five minutes, the full-sample coefficient is 1.610 percentage points with a market-clustered standard error of 0.151 and a wild-cluster  $p$ -value of 0.0001. Moving depth pressure from 0.25 to 2 implies 1.409 percentage points, or 1,409 dollars per 100,000 dollars of binary payout exposure. Strict same-token isolation retains 32,976 episodes from 1,804 markets and reduces the coefficient to 0.659 percentage points. The corresponding economic contrast is 577 dollars, with a 95 percent bootstrap- $t$  pivot interval from 298 to 854 dollars.

The horizon pattern distinguishes immediate movement from persistence. The isolated coefficient is 0.483 percentage points after 60 seconds and 0.548 after 15 minutes, with wild-cluster  $p$ -values of 0.0001 and 0.0011. After one hour, however, the isolated coefficient is 0.464 percentage points with a  $p$ -value of 0.1476. The full-sample coefficient remains positive at one hour, but only 5.0 percent of available one-hour outcomes are isolated. The 28-day sample therefore supports a short-run depth-pressure relation while providing no statistically detectable isolated one-hour effect.

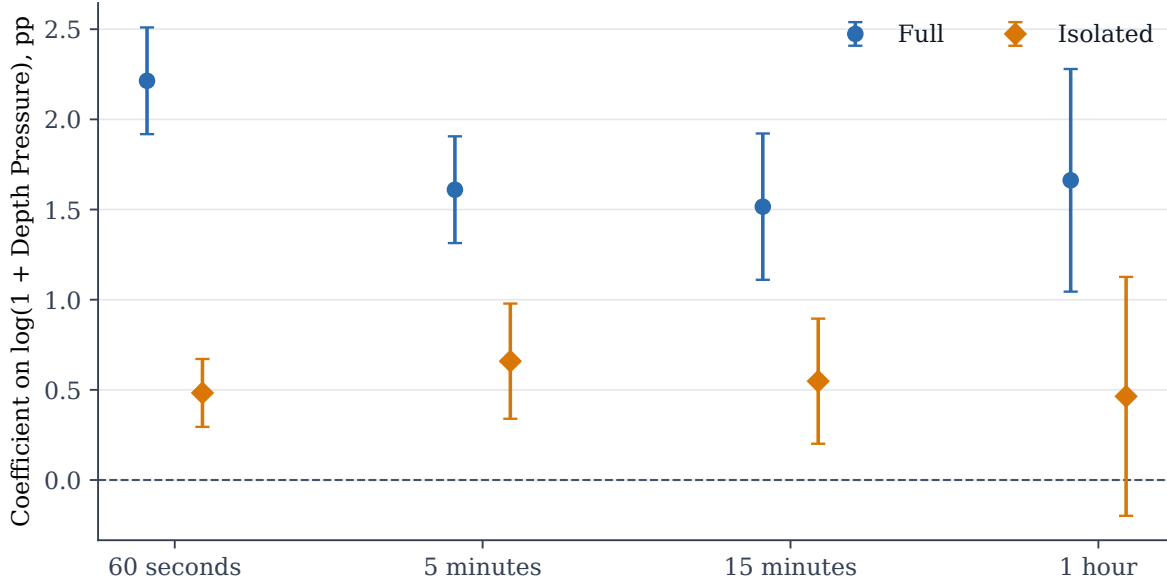


Figure 1: Depth pressure and directional price response across horizons

*Notes:* Points are coefficients on log one plus top-5 directional depth pressure in the 28-day prospective sample. Bars are normal 95 percent intervals based on market-clustered standard errors. Full-sample and strictly isolated estimates use the same specification. Full-sample episode counts from 60 seconds to one hour are 185,235, 173,240, 152,676, and 100,717. Isolated counts are 56,251, 32,976, 16,124, and 5,003.

Recorder coverage limits the scope of this evidence. Joint trade-and-orderbook coverage spans 20.71 days in 19 segments, and the largest gap lasts 2.06 days. Requiring the entire event window to lie within a continuously observed segment yields a five-minute coefficient of 1.658 percentage points. The longest continuous segment yields 1.452. Episodes within one hour of a coverage boundary yield 1.188, compared with 1.636 away from boundaries. These checks preserve the sign but show that coverage gaps can affect magnitude. Appendix Table A2 reports the coverage checks together with alternative thresholds, episode definitions, outlier treatments, and inference procedures. The sample should therefore be interpreted as a broad prospective collection from a rolling active-market universe, not as a continuous census of Polymarket.

## 4.2 Forecast Performance and Incremental Information

The central price-response estimates do not by themselves establish whether post-event prices are better forecasts or whether any improvement is specific to large orders. I therefore link a separate set of episodes to realized binary outcomes. The sample spans August 14, 2026 at 12:06:49 UTC through September 1, 2026 at 23:59:59 UTC and excludes every market used in the earlier samples.

Table 6: Forecast performance and incremental information in the later sample

|                                                                | Estimate  | Std. error | 95% interval       | p-value |
|----------------------------------------------------------------|-----------|------------|--------------------|---------|
| <i>Panel A: Prices around isolated large-order episodes</i>    |           |            |                    |         |
| Pre-episode Brier loss                                         | 0.1516    |            |                    |         |
| Five-minute Brier loss                                         | 0.1454    |            |                    |         |
| Post-minus-pre change                                          | -0.0062   | 0.0010     | [-0.0083, -0.0042] | 0.0001  |
| Fixed-bin miscalibration change                                | -0.0002   | 0.0004     | [-0.0009, 0.0005]  | 0.2745  |
| <i>Panel B: Large orders versus comparable ordinary trades</i> |           |            |                    |         |
| Large-order Brier change                                       | -0.0024   |            |                    |         |
| Ordinary-trade Brier change                                    | -0.0026   |            |                    |         |
| Large-minus-ordinary difference                                | 0.0001    | 0.0010     | [-0.0019, 0.0021]  | 0.9042  |
| <i>Panel C: Incremental out-of-time forecast performance</i>   |           |            |                    |         |
| Signed-flow model minus baseline                               | -0.000005 | 0.0002     | [-0.0004, 0.0004]  | 0.9838  |

Notes: Brier loss is  $(p - y)^2$ , where  $p$  is the traded token's price and  $y$  is its resolved binary payoff. Panel A reports market-equal values around 19,072 isolated large-order episodes in 1,706 markets. Its Brier  $p$ -value is from a one-sided restricted wild test, while the calibration  $p$ -value is from an event-family cluster bootstrap. Panel B compares the five-minute Brier change after 11,235 large-order episodes with that after 74,889 comparable ordinary trades in 1,367 markets. Overlap weights balance starting probability, time to settlement, recent activity, volatility, spread, and event time. Panel C compares predictions from a baseline model using the pre-episode price and unsigned depth pressure with an augmented model that adds signed depth pressure. Both models are estimated in the earlier sample and evaluated without refitting on 20,167 later episodes in 1,706 nonoverlapping markets. Panels B and C report event-family-clustered standard errors and two-sided restricted wild  $p$ -values based on 9,999 sign flips. The comparisons are retrospective and do not identify causal effects.

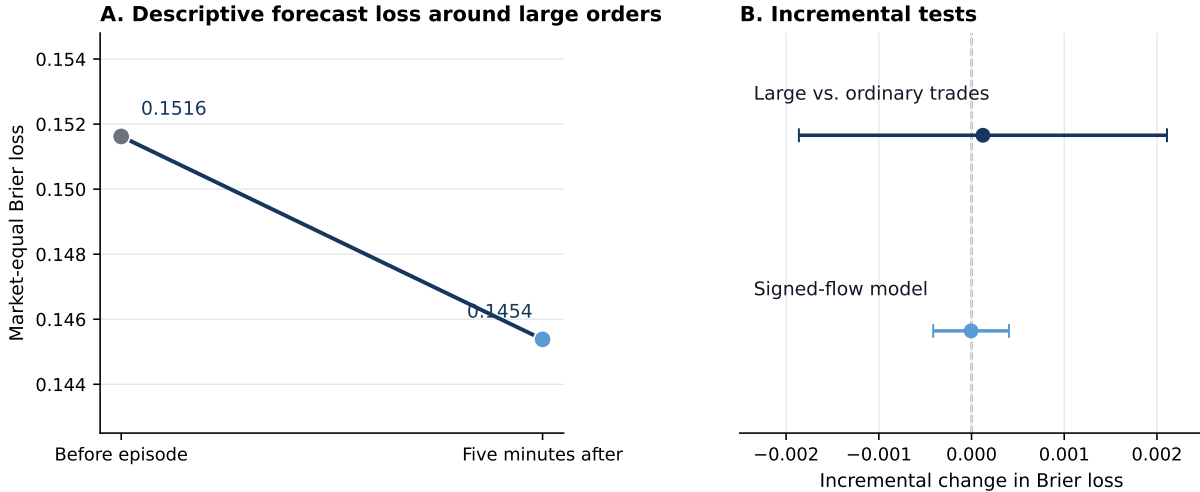


Figure 2: Forecast performance and incremental information

Notes: Panel A reports market-equal Brier loss immediately before and five minutes after isolated large-order episodes. The plotted values match Table 6 and are rounded to four decimals. Panel B reports two incremental comparisons with 95 percent confidence intervals. The first is the large-minus-ordinary difference in five-minute Brier changes from the overlap-weighted risk-set design. The second is the augmented-minus-baseline Brier loss from the out-of-time prediction exercise. Negative values indicate lower forecast loss for large orders or for the model that adds signed order flow. The two-sided restricted wild  $p$ -values are 0.9042 for the risk-set comparison and 0.9838 for the signed-flow comparison.

Table 6 and Figure 2 first show that market-equal Brier loss falls from 0.1516 before the episodes to 0.1454 five minutes afterward. The change is -0.0062, a 4.11 percent reduction, with an event-family-clustered standard error of 0.0010 and a one-sided wild-test  $p$ -value of 0.0001. In this realized-outcome sense, post-event prices are closer on average to terminal outcomes than pre-event prices. The calibration diagnostic declines by 0.0002, but its 95 percent interval includes zero and I do not detect a calibration improvement. Across alternative calendar, rolling, and expanding windows, the Brier changes are generally negative. These overlapping partitions were constructed after outcomes were observed and appear only as descriptive evidence in Appendix Figure A1.

The risk-set comparison changes the interpretation of this descriptive improvement. In the overlap-weighted sample, Brier loss declines by 0.0024 after large orders and by 0.0026 after comparable ordinary trades. The large-minus-ordinary difference is 0.0001, with a standard error of 0.0010, a 95 percent interval from -0.0019 to 0.0021, and a two-sided wild  $p$ -value of 0.9042. All specified support and balance thresholds are satisfied, including a maximum weighted standardized difference of 0.0227. I therefore do not detect a large-order-specific improvement in realized forecast loss. The common decline is consistent with information arriving over five minutes or with prices generally becoming more informative as events approach resolution.

The out-of-time prediction test reaches the same conclusion from a different direction. The baseline model’s market-equal Brier loss is 0.1421. Adding signed depth pressure changes the loss by -0.000005, with a standard error of 0.0002, a 95 percent interval from -0.0004 to 0.0004, and a two-sided wild  $p$ -value of 0.9838. I likewise do not detect a log-loss improvement. The interval bounds the average change tightly around zero in this evaluation sample. Because the trade direction is independently reconciled to the exchange events for 1,000 later-sample transactions, the null result is not readily attributed to systematic reversal of BUY and SELL labels. Together, the two tests show that the initial Brier decline is not sufficient evidence that large-order adjustment contains incremental outcome information.

### 4.3 Historical Order Books and Execution

The PMXT archive extends the depth analysis beyond the prospective day and provides message order within each archived hour. The sample covers 36 event hours and 144 surrounding hours with 5,671,153 filtered book events. It reconstructs 1,384 strict pre-trade books and matches every transaction hash over the complete episode for 1,317 observations from 30 markets and 913 wallets. A separate replay reproduces 5,238 of 5,268 bid, ask, and top-5 depth metrics, an agreement rate of 99.43 percent. This high agreement provides a measurement check on the reconstructed depth used in the historical analysis. The analysis retains 1,055 quote-eligible episodes after requiring strong agreement with direct book updates, positive relevant depth, and pre- and post-episode states no more than 60 seconds old.

Table 7 reports a dynamic decomposition on identical one-hour support. In the full sample, the observed P10-to-P90 pressure change predicts 858 dollars of immediate quote movement per 100,000 dollars of exposure and 1,753 dollars of subsequent reversal. The immediate coefficient

Table 7: Historical exact-orderbook movement and one-hour reversal

| Sample          | Outcome                  | $N$ | Markets | Coef.   | pp    | Wild $p$ | Dollars |
|-----------------|--------------------------|-----|---------|---------|-------|----------|---------|
| Full            | Immediate movement       | 317 | 24      | 12.478  | 0.066 |          | 858     |
| Full            | Reversal at one hour     | 317 | 24      | 25.506  | 0.029 |          | 1,753   |
| Full            | Total change at one hour | 317 | 24      | -13.028 | 0.186 |          | -896    |
| Selected hours  | Immediate movement       | 148 | 21      | 12.790  | 0.010 |          | 2,249   |
| Selected hours  | Reversal at one hour     | 148 | 21      | 24.206  | 0.003 |          | 4,257   |
| Selected hours  | Total change at one hour | 148 | 21      | -11.416 | 0.259 |          | -2,008  |
| Sport only      | Immediate movement       | 114 | 12      | 13.831  | 0.001 |          | 1,280   |
| Sport only      | Reversal at one hour     | 114 | 12      | 25.196  | 0.005 |          | 2,331   |
| Sport only      | Total change at one hour | 114 | 12      | -11.365 | 0.333 |          | -1,051  |
| Excluding sport | Immediate movement       | 34  | 9       | -0.299  | 0.312 |          | -118    |
| Excluding sport | Reversal at one hour     | 34  | 9       | -0.849  | 0.750 |          | -336    |
| Excluding sport | Total change at one hour | 34  | 9       | 0.549   | 0.766 |          | 217     |

Notes: Coef. pp is the coefficient on log one plus top-5 depth pressure in percentage points. Models control for log episode notional, the pre-trade spread, the directional 60-second pre-return, trade side, market fixed effects, UTC-hour fixed effects, and log one plus the exact 24-hour pre-event Google News count. Standard errors are clustered by market. Wild  $p$  uses 9,999 market-level Rademacher replications. Immediate movement is the change in the execution-side quote from the strict pre-trade state to the state after the final episode trade. Reversal is the immediate quote minus the one-hour quote, so a positive value indicates retracement. Total change is the one-hour quote minus the pre-trade quote. Every row uses common one-hour support. Dollars scale each coefficient to the P10-to-P90 pressure change observed within that row's sample and report value per 100,000 dollars of binary payout exposure. Sport-only rows use the outcome-independent selection. The non-sport rows are identical in the full and selected samples because every stress hour is a sport market.

is marginal under wild-cluster inference with  $p = 0.066$ . The reversal coefficient has  $p = 0.029$ , while the one-hour total change is statistically insignificant. In the 32 outcome-independent event hours, the corresponding immediate movement is 2,249 dollars and the reversal is 4,257 dollars, with wild-cluster  $p$ -values of 0.010 and 0.003. Exact 60-second and five-minute pre-trend placebos are insignificant under wild-cluster inference. Adding the pre-event news count changes neither coefficient economically.

The market split materially narrows the interpretation. In selected sports markets, the P10-to-P90 immediate movement is 1,280 dollars and the reversal is 2,331 dollars. Outside sports, the immediate movement is minus 118 dollars and the reversal is minus 336 dollars, with wild-cluster  $p$ -values of 0.312 and 0.750. The historical design therefore documents a temporary quote response in selected event-proximate sports markets, not a platform-wide large-order effect. Google News controls do not absorb this relation, but headline counts do not measure live scores or scheduled game progression. Public information remains a plausible common cause of large orders and quote changes.

The reconstructed books also permit an execution-level book walk. Among 833 single-trade, non-exhausted episodes from 29 markets, actual entry slippage averages 0.144 percentage points and the static book predicts 0.174 percentage points. The mean absolute prediction error is 0.150 percentage points, or 150 dollars per 100,000 dollars of exposure. Predicted and actual slippage have a Spearman correlation of 0.544. The calibration slope is 0.630 and differs from zero with

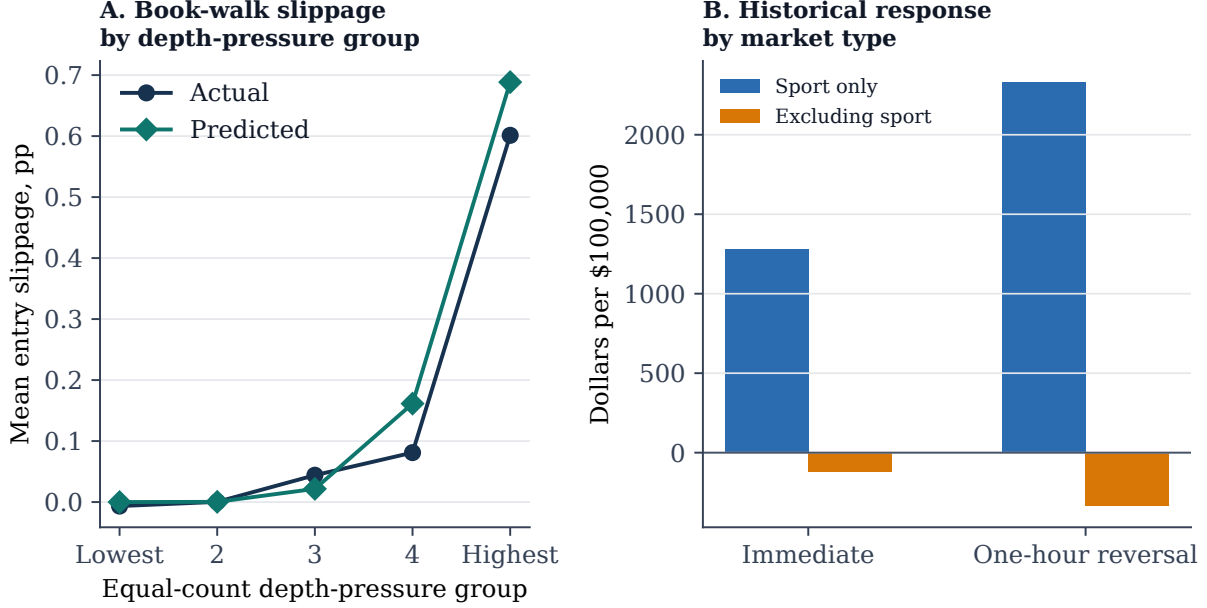


Figure 3: Exact order books, execution costs, and short-run reversal

*Notes:* Panel A sorts the single-trade, non-exhausted book-walk sample into five equal-count groups by pre-trade top-5 depth pressure and compares mean actual entry slippage with the static book-walk prediction. The groups contain 166, 167, 166, 167, and 167 episodes. These group means are descriptive and do not carry separate confidence intervals. Panel B reports P10-to-P90 depth-pressure contrasts in dollars per 100,000 dollars of binary payout exposure on common one-hour support. Sport-only estimates use 114 episodes in 12 markets, with two-sided wild-cluster  $p$ -values of 0.001 for immediate movement and 0.005 for reversal. Non-sport estimates use 34 episodes in nine markets, with  $p$ -values of 0.312 and 0.750.

a wild-cluster  $p$ -value of 0.001. The evidence shows that book-implied slippage ranks realized execution costs. The slope is below one, but the corresponding test against a unit slope is only suggestive ( $p = 0.069$ ), so systematic overprediction remains uncertain.

Table 8 shows why directional price impact is not equivalent to trader profit. Across all prospective whale episodes, the five-minute midpoint moves by 26 dollars per 100,000 dollars of payout exposure in the trade direction. The volume-weighted episode price is already 90 dollars worse than the pre-trade midpoint. Marking the position to the five-minute midpoint therefore produces a loss of 64 dollars. Closing at the best quote increases the loss to 133 dollars. The large-trade-only sample has a best-quote loss of 170 dollars, with a market-clustered 95 percent interval from 254 dollars of losses to 87 dollars of losses. Only 13.6 percent of these episodes have positive best-quote profit after five minutes.

The high-pressure episodes provide the sharpest decomposition. Episodes above twice top-5 depth are followed by a midpoint change of 864 dollars per 100,000 dollars of exposure. Their episode price is 866 dollars worse than the pre-trade midpoint. The two components nearly offset, leaving a midpoint loss of 3 dollars and a best-quote loss of 105 dollars. Among episodes that also satisfy the large-trade definition, the quote change equals 663 dollars while the entry mark equals minus 808 dollars. Their best-quote loss is 246 dollars. High-pressure episodes are therefore

Table 8: Five-minute profitability from the actual episode price

| Sample                         | Episodes | Price impact | Entry mark | Midpoint P&L | Best-quote P&L | Positive (%) |
|--------------------------------|----------|--------------|------------|--------------|----------------|--------------|
| All whale episodes             | 4,051    | 25.7         | -89.6      | -63.9        | -133.2         | 11.7         |
| Large-trade only               | 1,466    | 18.3         | -116.4     | -98.0        | -170.3         | 13.6         |
| Depth pressure $\leq 0.25$     | 3,645    | 8.5          | -64.7      | -56.2        | -121.9         | 8.9          |
| Depth pressure $> 2$           | 61       | 863.9        | -866.5     | -2.6         | -105.0         | 36.1         |
| Large trade and pressure $> 2$ | 53       | 663.2        | -807.9     | -144.7       | -245.6         | 35.8         |
| Isolated episodes              | 958      | 0.0          | -60.5      | -60.5        | -125.2         | 3.1          |

Notes: Values are dollars per 100,000 dollars of binary payout exposure except for the episode count and positive share. Price impact is the directional movement from the pre-episode midpoint to the five-minute midpoint. Entry mark is the directional difference between the pre-episode midpoint and the volume-weighted episode price. Midpoint P&L starts from the episode price. Best-quote P&L closes Buy episodes at the future best bid and Sell episodes at the future best ask. The positive column reports the percentage of best-quote profits above zero. Measures are gross of platform and gas fees. Sample counts can differ slightly across columns because quote availability differs.

associated with a movement in the public quote, but the traders initiating those orders pay for most or all of that movement through their execution price.

Best-quote profit assumes that the entire position can exit at the displayed quote. Restricting the large-trade sample to 1,007 five-minute observations with enough top-of-book quantity still yields an average loss of 122 dollars per 100,000 dollars of payout exposure. Only 6.2 percent are profitable. Strictly isolated large-trade episodes lose 129 dollars at the best quote. The result is also negative for Buy and Sell episodes separately.

Ordinary trades provide a second benchmark. Their average five-minute best-quote loss is 138 dollars per 100,000 dollars of payout exposure. Large trades lose 156 dollars at the trade level. In the preferred regression with token and hour fixed effects, trade-direction controls, and entry-price controls, the adjusted large-trade difference is minus 11 dollars. Its market-clustered 95 percent interval ranges from minus 95 to plus 73 dollars. At one hour, the adjusted difference is 27 dollars with an interval from minus 180 to plus 234 dollars. The data therefore do not show economically large or statistically robust short-horizon excess profitability for large trades. Appendix Table A6 shows the corresponding comparisons across horizons and alternative marking rules. Final outcomes remain unobserved for this active-market panel, so the separate resolved-market sample provides the settlement comparison.

The resolved-market sample complements this short-horizon result with final payouts. Table 9 reports the preferred within-market 95th-percentile definition. The 30-market out-of-sample estimate compares whales and nonwhales in markets selected before their fill outcomes were analyzed. The combined estimate adds the 20 highest-volume eligible markets and provides the most precise magnitude.

In the out-of-sample markets, the overall whale-minus-nonwhale difference is 1,043 dollars per 100,000 dollars of exposure, but its interval from minus 189 to 2,274 dollars includes zero. The role decomposition is more informative. Whale takers earn 1,669 dollars per 100,000 dollars, while nonwhale takers lose 1,246 dollars. Their market-equal difference is 2,642 dollars, with an interval

Table 9: Realized whale profitability in resolved markets

| Sample           | Role  | Whale P&L | Nonwhale P&L | Difference | 95% interval    | Positive markets |
|------------------|-------|-----------|--------------|------------|-----------------|------------------|
| Out-of-sample 30 | All   | 239       | -925         | 1,043      | [-189, 2,274]   | 24/30            |
|                  | Maker | -967      | -292         | -1,025     | [-2,960, 909]   | 19/30            |
|                  | Taker | 1,669     | -1,246       | 2,642      | [525, 4,760]    | 25/30            |
| Combined 50      | All   | 216       | -835         | 1,075      | [317, 1,832]    | 42/50            |
|                  | Maker | 433       | -261         | -107       | [-1,322, 1,108] | 37/50            |
|                  | Taker | -53       | -1,071       | 1,829      | [546, 3,111]    | 40/50            |

Notes: P&L and differences are dollars per 100,000 dollars of binary payout exposure. The preferred whale definition classifies a fill when either its notional or its wallet-market cumulative notional is at or above the within-market 95th percentile. P&L levels aggregate exposure within each sample. Differences and Student-*t* intervals are market-equal. Maker denotes passive orders and taker denotes aggressive orders. Recorded exchange fees are included. Gas costs are excluded.

from 525 to 4,760 dollars. Whale makers lose 967 dollars and do not outperform nonwhale makers. The aggressive side therefore generates the relative advantage.

The combined 50-market sample yields a market-equal overall difference of 1,075 dollars, with an interval from 317 to 1,832 dollars. The taker difference is 1,829 dollars, with an interval from 546 to 3,111 dollars, while the maker difference is minus 107 dollars and statistically indistinguishable from zero. The absolute and relative magnitudes imply different economic conclusions. Combined-sample whale takers earn minus 53 dollars per 100,000 dollars of exposure. Their advantage arises because nonwhale takers lose 1,071 dollars, not because whales earn unusually high absolute returns. The evidence is therefore more consistent with superior execution, order selection, or avoidance of adverse trades than with systematic information rents.

The resolved-market estimates complement the short-horizon evidence. Large aggressive orders are followed by changes in the public quote, and large takers perform better than smaller takers relative to final payouts. Neither the five-minute marks nor combined absolute settlement returns show exceptional average profits. Appendix Tables A7 and A8 show how the settlement comparison varies across whale definitions and pre-specified market-volume strata. The average magnitudes do not support a broad insider-trading interpretation.

The retrospective classification uses future market activity and may mechanically identify wallets that become large later. Table 10 removes this look-ahead. The initial-period design estimates q95 thresholds from the first 20 percent of distinct trading timestamps and applies them to the remaining observations. The expanding design recalculates thresholds before each evaluation block using only earlier blocks. The last two rows in each panel isolate prior wallet activity within the same market and across earlier non-overlapping markets.

The expanding union estimate confirms the role-specific result without future information. In the 30 out-of-sample markets, large taker orders outperform smaller taker orders by 2,358 dollars per 100,000 dollars of payout exposure. The interval is 292 to 4,423 dollars and the wild-sign *p*-value is 0.018. The corresponding combined-sample difference is 1,493 dollars, with an interval from 234 to 2,751 dollars and a wild-sign *p*-value of 0.012. The whale-by-taker interaction equals

Table 10: Ex-ante whale classification and resolved-market profitability

| Sample      | Classification           | Taker diff. | 95% interval    | Wild $p$ | Interaction | 95% interval    | Wild $p$ |
|-------------|--------------------------|-------------|-----------------|----------|-------------|-----------------|----------|
| OOS 30      | Initial 20 percent union | 1,661       | [-527, 3,848]   | 0.136    | 3,929       | [87, 7,771]     | 0.033    |
|             | Expanding union          | 2,358       | [292, 4,423]    | 0.018    | 4,276       | [710, 7,843]    | 0.005    |
|             | Expanding fill size      | 2,496       | [118, 4,874]    | 0.037    | 5,236       | [1,218, 9,253]  | 0.004    |
|             | Expanding prior wallet   | 1,278       | [-1,577, 4,133] | 0.416    | -489        | [-3,006, 2,028] | 0.721    |
|             | Prior-market wallet      | 952         | [-2,695, 4,600] | 0.636    | 573         | [-3,180, 4,327] | 0.760    |
| Combined 50 | Initial 20 percent union | 1,044       | [-267, 2,356]   | 0.120    | 2,136       | [-223, 4,495]   | 0.087    |
|             | Expanding union          | 1,493       | [234, 2,751]    | 0.012    | 2,373       | [167, 4,579]    | 0.021    |
|             | Expanding fill size      | 1,525       | [78, 2,972]     | 0.035    | 2,868       | [343, 5,392]    | 0.018    |
|             | Expanding prior wallet   | 1,197       | [-515, 2,909]   | 0.181    | -76         | [-1,595, 1,442] | 0.920    |
|             | Prior-market wallet      | 585         | [-1,950, 3,119] | 0.668    | 514         | [-2,013, 3,042] | 0.693    |

Notes: Taker difference is whale taker P&L minus nonwhale taker P&L. Interaction subtracts the corresponding maker difference. All magnitudes are dollars per 100,000 dollars of binary payout exposure and are averaged equally across markets. Initial-period thresholds use the first 20 percent of distinct trading timestamps. Expanding thresholds use only earlier timestamp blocks. Order size uses the current order size relative to the lagged distribution. Prior wallet uses only cumulative wallet activity observed before the evaluation block. Prior-market wallet uses non-overlapping earlier market histories. Wild  $p$ -values use 9,999 market-level Rademacher sign flips.

4,276 dollars out of sample and 2,373 dollars in the combined sample. Both intervals exclude zero. These estimates show that the relative settlement advantage is concentrated on the aggressive side and is not shared by whale makers.

The initial-period design produces smaller and less precise direct taker differences. Its combined difference is 1,044 dollars with an interval from minus 267 to 2,356 dollars. The out-of-sample difference is 1,661 dollars with an interval from minus 527 to 3,848 dollars. The corresponding out-of-sample interaction remains positive at 3,929 dollars, with an interval from 87 to 7,771 dollars. The similarity in point estimates and the loss of precision are consistent with discarding the initial estimation period and fixing noisier early thresholds. These results support the direction of the expanding estimates but do not independently establish a direct taker difference at conventional levels.

The components distinguish order-specific performance from persistent trader identity. Under the expanding fill-only classification, the taker difference is 2,496 dollars out of sample and 1,525 dollars in the combined sample. Both intervals exclude zero, and the combined whale-by-taker interaction is 2,868 dollars with an interval from 343 to 5,392 dollars. In contrast, the expanding prior-wallet interaction is minus 76 dollars in the combined sample, with an interval from minus 1,595 to 1,442 dollars. A wallet label constructed from earlier non-overlapping markets yields a taker difference of 585 dollars with an interval from minus 1,950 to 3,119 dollars. Previously observed wallet scale therefore does not predict persistent performance. Appendix Tables A9 and A10 report sensitivity to alternative ex-ante thresholds and variation across the market life cycle.

Absolute profit differs across cohorts. Combined-sample large takers earn minus 48 dollars per 100,000 dollars under the expanding union definition, while smaller takers lose 803 dollars in the exposure-weighted calculation. A market-cluster bootstrap gives a 90 percent interval from minus 539 to 564 dollars. The two one-sided tests reject returns outside a 1,000 dollar equivalence margin

Table 11: Large Kalshi trades, local prices, and matched Polymarket prices

| Outcome                | Events | Difference | 95% interval    | <i>p</i> -value |
|------------------------|--------|------------|-----------------|-----------------|
| Settlement P&L         | 33     | 365        | [-1,460, 2,191] | 0.708           |
| Kalshi, 5 minutes      | 32     | 9          | [-1,032, 1,050] | 0.974           |
| Kalshi, 30 minutes     | 31     | 300        | [-111, 711]     | 0.172           |
| Kalshi, 6 hours        | 29     | -179       | [-640, 282]     | 0.458           |
| Polymarket, 5 minutes  | 33     | 192        | [-106, 489]     | 0.239           |
| Polymarket, 30 minutes | 33     | 448        | [17, 879]       | 0.031           |
| Polymarket, 6 hours    | 31     | 154        | [-209, 516]     | 0.421           |

Notes: Large trades are at or above the q95 contract-count threshold estimated from the first 20 percent of each Kalshi market’s distinct trading timestamps. Calibration trades are excluded. Differences compare large with other trades and are dollars per 100,000 dollars of binary exposure. Events are conservative macro-event clusters. Intervals use Student-*t* critical values. The *p*-values use 9,999 event-level Rademacher sign flips. Local and cross-platform prices must fall within ten minutes of the target horizon. Settlement P&L is gross of Kalshi fees.

with a joint *p*-value of 0.002. Appendix Table A14 reports the equivalence tests and minimum detectable effects. The out-of-sample large-taker return is 1,835 dollars, with a 90 percent interval from 478 to 3,257 dollars. This positive level is concentrated. Three sports markets generate 72.6 percent of aggregate out-of-sample large-taker profit. The return falls to 599 dollars after excluding these markets, 262 dollars outside sports, and minus 62 dollars more than seven days before market close. The corresponding 95 percent market-cluster bootstrap intervals include zero. Appendix Table A11 reports these concentration and timing diagnostics. Price filters leave the positive level at least as strong, which rules out mechanically extreme contract prices as its source. The combined evidence supports an order-size channel involving execution, timing, or short-horizon trade selection, but not a uniform information advantage across cohorts or market categories.

The cross-platform sample tests whether the order-size pattern extends beyond Polymarket. Table 11 reports the q95 estimates. Large Kalshi trades outperform other Kalshi trades at settlement by 365 dollars per 100,000 dollars of payout exposure. The 95 percent interval ranges from minus 1,460 to 2,191 dollars and the randomization *p*-value is 0.708. The approximate 80 percent minimum detectable effect is 2,610 dollars. Equivalence tests reject differences outside a 2,000 dollar margin, but not outside a 1,000 dollar margin. The absolute large-trade return has a 90 percent interval from minus 2,183 to 4,365 dollars and is not equivalent to zero within either margin. The evidence therefore rejects an economically very large relative advantage under q95, but it does not rule out moderate settlement alpha or a large absolute return with adequate precision.

Local Kalshi transaction prices provide limited precision under the strict price-distance rule. The q95 difference is 9 dollars after five minutes and 300 dollars after 30 minutes. Both intervals include zero. After six hours, the estimate is minus 179 dollars. The matched Polymarket response is more concentrated at 30 minutes. Large Kalshi trades predict a 448 dollar movement per 100,000 dollars in the same direction, equivalent to 0.448 percentage points in the implied event probability. The interval is 17 to 879 dollars and the randomization *p*-value is 0.031. The five-minute estimate is 192 dollars and the six-hour estimate is 154 dollars. Neither is statistically distinguishable from zero.

Table 12: Descriptive Google News RSS audit for top 100 whale episodes

| News bin  | Episodes | Mean articles | 60s pp | 1h pp |
|-----------|----------|---------------|--------|-------|
| No news   | 25       | 0.0           | 0.950  | 1.242 |
| Low news  | 21       | 3.5           | 0.358  | 0.380 |
| High news | 54       | 49.5          | 0.471  | 0.624 |

The cross-platform pattern rises between five and 30 minutes and becomes imprecise by six hours. Appendix Table A13 separates prior price movement, source-price adjustment, and sample selection. The 30-minute Polymarket lead placebo is 262 dollars per 100,000 dollars, with a randomization  $p$ -value of 0.673. In the 246-market lead-available sample, the raw 30-minute response is 448 dollars. Controlling for the prior Polymarket movement reduces the estimate to 374 dollars. Its randomization  $p$ -value is 0.047 before and 0.140 after Holm correction across the three post-trade horizons.

The complete source-price specification requires both prior and future Kalshi transaction prices within ten minutes of their targets. This restriction retains 109 markets and 27 macro-events at 30 minutes. The raw estimate in this common sample is 431 dollars and is already imprecise. Adding the prior Polymarket and Kalshi movements produces 517 dollars. Adding the same-horizon Kalshi return reduces the conditional response to 130 dollars, with a 95 percent interval from minus 222 to 482 dollars and a randomization  $p$ -value of 0.412. The corresponding q90 and q99 estimates are 155 and 176 dollars and are also insignificant. Because the same-horizon Kalshi return is measured after the trade, it is a transmission variable and not an exogenous control. I do not detect an incremental cross-venue signal in large trade size once the source venue’s own price adjustment is held fixed.

Unadjusted threshold and event-family checks support a descriptive 30-minute lead-lag relation. Appendix Table A12 reports these specifications. The q90, q95, and q99 estimates equal 514, 448, and 480 dollars per 100,000 dollars, with randomization  $p$ -values of 0.006, 0.031, and 0.012. Across 15 leave-one-family-out specifications, the unadjusted estimate remains positive between 284 and 645 dollars. The full-control estimates instead equal 155, 130, and 176 dollars under q90, q95, and q99. Their Holm-adjusted randomization  $p$ -values are 1.000, 0.823, and 0.933. The raw relation is therefore robust as cross-venue predictability, but not as an order-size effect independent of Kalshi’s contemporaneous price adjustment. Settlement P&L remains insignificant under every threshold and family exclusion.

The analysis queries Google News RSS for 100 episodes selected after observing large event-study movements, using market-specific keywords and a one-day window on each side of the episode. The returned titles support manual relevance and direction coding. Because selection depends on realized price movement, the exercise is descriptive.

Table 12 shows that the selected large movements are not confined to news-heavy windows. Twenty-five of the top 100 episodes have no Google News articles in the event window. Their mean one-hour directional movement is 1.242 percentage points, compared with 0.624 percentage points

Table 13: Manual news coding for top 100 whale episodes

| Manual code                  | Episodes | Mean articles | 60s pp | 1h pp |
|------------------------------|----------|---------------|--------|-------|
| No relevant news             | 25       | 0.0           | 0.950  | 1.242 |
| Irrelevant or noisy news     | 7        | 9.6           | 0.667  | 0.667 |
| Ambiguous relevant news      | 27       | 35.7          | 0.268  | 0.723 |
| News against whale direction | 15       | 31.5          | 0.635  | 0.878 |
| News in whale direction      | 26       | 47.8          | 0.427  | 0.262 |

Table 14: Regression-adjusted manual news effects

| Manual code                  | Episodes | Raw pp | Adjusted pp | Dollars |
|------------------------------|----------|--------|-------------|---------|
| News in whale direction      | 13       | 0.262  | 0.252       | 252     |
| No relevant news             | 13       | 1.242  | 1.498       | 1,498   |
| Irrelevant or noisy news     | 3        | 0.667  | 0.872       | 872     |
| Ambiguous relevant news      | 11       | 0.723  | 0.454       | 454     |
| News against whale direction | 9        | 0.878  | 0.782       | 782     |

for episodes with high news intensity. The RSS procedure therefore identifies no contemporaneous article for some economically large movements, although keyword matching cannot establish that no public information existed. A 1.242 percentage point movement changes the marked value of 100,000 dollars of exposure by 1,242 dollars.

The top 100 RSS episodes are manually coded using the article titles returned by the news query.<sup>3</sup>

Table 13 refines the descriptive news audit. Twenty-five of the top 100 episodes contain no relevant news in the RSS window and move by an average of 1.242 percentage points after one hour, or about 1,242 dollars per 100,000 dollars of binary payout exposure. The seven episodes coded as irrelevant or noisy news move by 0.667 percentage points, or about 667 dollars per 100,000 dollars of exposure, while episodes with news in the whale direction move by 0.262 percentage points. Clearly directional public news is therefore not present in every selected episode, but these post-outcome groups do not estimate population news effects.

Table 14 reports a regression-adjusted description for episodes with complete one-hour price windows. The regression includes log episode notional and the number of returned RSS articles, with the news-in-whale-direction group as the omitted category. The adjusted mean for episodes without relevant news is 1.498 percentage points, or about 1,498 dollars per 100,000 dollars of exposure. The adjusted mean for irrelevant or noisy news episodes is 0.872 percentage points. With 49 complete one-hour observations selected after the outcome, these estimates remain descriptive. Adjustment for article volume and episode size does not reduce the no-news mean toward zero.

Matched same-token timestamps provide a descriptive benchmark for ordinary movement at similar token prices. Table 15 compares whale episodes with complete one-hour windows to times-

<sup>3</sup>The codebook separates no relevant news, irrelevant or noisy news, ambiguous relevant news, news in the whale direction, and news against the whale direction. A title is directional only when it contains information that plausibly shifts the contract probability, not merely a candidate, officeholder, or institution mention.

Table 15: Matched same-token timestamp benchmark

| Horizon    | Whale mean | Control mean | Difference |
|------------|------------|--------------|------------|
| 60 seconds | 0.0010     | 0.0000       | 0.0009     |
| 5 minutes  | 0.0012     | 0.0000       | 0.0011     |
| 15 minutes | 0.0012     | 0.0001       | 0.0012     |
| 1 hour     | 0.0011     | 0.0001       | 0.0011     |

Table 16: Bootstrap confidence intervals for whale-control differences

| Horizon    | Difference | 2.5 percent | 97.5 percent |
|------------|------------|-------------|--------------|
| 60 seconds | 0.0009     | 0.0004      | 0.0016       |
| 5 minutes  | 0.0011     | 0.0005      | 0.0019       |
| 15 minutes | 0.0012     | 0.0005      | 0.0020       |
| 1 hour     | 0.0011     | 0.0004      | 0.0018       |

tamps in the same token, matched on base price and placed away from whale windows. At 60 seconds, the whale mean is 0.095 percentage points and the assigned-control mean is 0.002 percentage points. The difference is 0.093 percentage points. The corresponding differences are 0.113 percentage points after five minutes, 0.117 after 15 minutes, and 0.108 after one hour. The 1,910 assigned control rows represent only 614 unique token-time pairs, so this benchmark does not provide 1,910 independent counterfactual observations.

The episode-resampling intervals are positive. Table 16 reports an interval of 0.043 to 0.160 percentage points for the 60-second whale-control difference. The intervals are 0.054 to 0.190 percentage points after five minutes, 0.051 to 0.203 after 15 minutes, and 0.044 to 0.182 after one hour. These contrasts correspond to 43 to 203 dollars per 100,000 dollars of exposure. Reused control timestamps create dependence that this bootstrap does not fully remove. The estimates therefore show that the selected whale timestamps differ from the matched benchmark, not that whale order flow causally generates the entire difference.

Across 16 matching specifications, the whale-control difference remains positive at all horizons from 60 seconds to one hour. Appendix Table A15 varies the number of controls and the event-exclusion window. The smallest one-hour difference is about 0.070 percentage points, while most one-hour specifications lie around 0.10 to 0.12 percentage points. The descriptive contrast is therefore not specific to one choice of control count or event-exclusion window. The grid does not address unobserved differences in information arrival between whale and matched timestamps.

Pre-trend diagnostics address whether whale episodes merely follow an existing directional movement. Table 17 shows pre-event means close to zero or negative over 60 seconds, five minutes, and 15 minutes, followed by positive post-event means. Appendix Figure A6 plots the same comparison across horizons. The pattern is inconsistent with simple short-horizon trend continuation. It does not distinguish mechanical price pressure from information arriving at the episode timestamp, because an informed trade can begin the observed movement.

Table 17: Pre-trend test

| Window     | Pre mean | Post mean | Post minus pre | 95 percent interval |
|------------|----------|-----------|----------------|---------------------|
| 60 seconds | -0.0002  | 0.0008    | 0.0010         | [0.0006, 0.0016]    |
| 5 minutes  | -0.0002  | 0.0010    | 0.0011         | [0.0006, 0.0018]    |
| 15 minutes | -0.0001  | 0.0010    | 0.0011         | [0.0005, 0.0019]    |
| 1 hour     | 0.0004   | 0.0012    | 0.0008         | [0.0000, 0.0016]    |

In both prospective samples, price movement increases with episode size relative to displayed pre-trade depth after market and hour fixed effects, large-trade restrictions, outlier treatment, leave-one-market-out estimation, strict event isolation, and restricted wild-cluster inference. The 28-day sample extends the relation to 2,271 markets. Alternative candidate thresholds, outlier treatments, episode windows, and continuous-coverage restrictions preserve its positive five-minute coefficient. The isolated one-hour estimate remains statistically indistinguishable from zero. The exact historical analysis adds a direct movement-and-reversal test and locates the one-hour round trip in sports, not across market categories.

The event-time orderbook results convert depth pressure into a market-design statistic. In the prospective 24-hour sample, the pressure increase from 0.25 to 2 implies 359 dollars of five-minute movement per 100,000 dollars of exposure and 214 dollars under strict isolation. In the 28-day sample, the same contrast implies 1,409 dollars in the full sample and 577 dollars under strict isolation. The larger estimate may reflect the rolling active-market composition and discontinuous coverage. Specifications restricted to continuously observed segments remain positive but vary in magnitude. The historical pressure distribution is much lower, so applying the same standardized contrast would extrapolate beyond common support. Its observed P10-to-P90 contrast instead implies 2,249 dollars of immediate movement and 4,257 dollars of reversal in the selected one-hour sample, but no economically or statistically robust relation outside sports.

The market-quality result concerns price pressure, not trader intent. Concentrated order flow and public-signal movement covary with available depth. The exact replay locates temporary movement in event-proximate sports, where live information can jointly determine order size and prices. The evidence does not establish a general causal pressure effect, but it shows that traded probabilities can move materially when displayed liquidity is thin and the information environment changes quickly.

Settlement outcomes distinguish order-specific performance from persistent trader skill. Orders that are large relative to the earlier size distribution perform better on the aggressive side, while lagged wallet scale does not predict settlement performance within or across markets. In the combined Polymarket sample, the absolute large-taker return is statistically equivalent to zero within 1,000 dollars per 100,000 dollars of exposure. The positive held-out return is concentrated in sports and in the final week before market close. Excluding sports leaves 262 dollars per 100,000 dollars, while trades more than seven days before close earn minus 62 dollars, and both market-cluster bootstrap intervals include zero. This timing points to event-proximate selection and execution

differences. The evidence weighs more strongly against a stable identity-based advantage than against episode-specific information.

## 5 Conclusion

Prediction-market prices are forecasts and transaction prices formed in finite-depth order books. Across two prospective samples, large-order episodes are followed by larger short-run quote changes when order size is high relative to displayed depth. Strict isolation weakens this association, and I do not detect a persistent isolated response after one hour. In the historical sample, book-implied slippage ranks realized execution costs, while reversal is concentrated in selected event-proximate sports markets.

The outcome tests sharpen the interpretation since Prices are closer to eventual outcomes five minutes after large-order episodes, but comparable ordinary trades improve almost as much. Adding signed order flow to the prevailing price and unsigned depth pressure produces no detectable out-of-time forecast improvement. A large directional trade is therefore not a sufficient signal of superior information, and its direction provides no detectable average forecasting advantage in the later sample.

The findings connect information aggregation with market microstructure by distinguishing price adjustment from incremental outcome information. Large orders are associated with larger short-run quote changes when displayed depth is low, but I do not detect settlement information in their direction beyond the prevailing price and unsigned depth pressure. Interpreting a prediction-market quote therefore requires considering both its realized forecast performance and the liquidity and trading process that produced it. This distinction reconciles substantial quote movement with weak average profitability and reversals in selected event-proximate sports markets.

The analysis does not observe the latent event probability or identify causal effects of liquidity, private information, trader intent, or manipulation. The samples emphasize active markets, the prospective collection contains gaps, and the historical evidence comes from selected hours. Live event information is imperfectly measured, especially in sports. Because the outcome tests use eventual settlements, unobserved information arrival may still differ between large and ordinary trades. Replication code and nonproprietary derived data will be public, but prospective order-book states are not fully reconstructible.

The implications are practical: Traders, risk managers, journalists, and policymakers should interpret a headline probability together with the liquidity supporting it, especially after a concentrated order. Platforms could display depth, spreads, recent volume, and large-transaction timing alongside quoted probabilities. The results neither identify individual trades as manipulation nor make short-run price movement evidence of private information. Prediction markets compress dispersed beliefs into a simple number. Interpreting it also requires understanding the trading process behind it.

## A Robustness Checks

### A.1 Trade-Direction Measurement

The signed-flow analysis requires a consistent mapping from each observed transaction to the taker’s direction in the traded outcome token. I validate this mapping against the exchange events recorded on Polygon for 1,000 transactions drawn from the later sample. A BUY denotes acquisition of the outcome token, while a SELL denotes disposal of that token. These labels do not denote Yes and No outcomes. A purchase of a No token is therefore a BUY in transaction terms even though its event exposure is negative.

Table A1: Trade-direction validation against blockchain events

|                                | Data interface | Blockchain   | Agreement |
|--------------------------------|----------------|--------------|-----------|
| Transactions                   | 1,000          | 1,000        | 100.0%    |
| BUY transactions               | 829            | 829          | 100.0%    |
| SELL transactions              | 171            | 171          | 100.0%    |
| Token quantity                 | 457,568.2049   | 457,568.2049 | 100.0%    |
| Unknown or ambiguous direction | 0              | 0            |           |

*Notes:* The table compares trades returned by Polymarket’s data interface with the corresponding exchange events recorded on Polygon. A match requires the same transaction hash, wallet, outcome token, taker direction, and token quantity. The exchange event identifies the taker’s order and records direction as BUY or SELL in the traded token. The validation sample contains 1,000 transactions from 646 market conditions. The total collateral notional is 211,732.5332 dollars in the data interface and 211,732.5332 dollars on-chain after rounding to four decimal places. Official contract documentation and source code are available through Polymarket’s contract documentation.

The transaction-level agreement supports the direction measure used in the signed-flow test. It does not independently validate beneficial ownership, informed trading, or the economic independence of wallets. Those questions require additional identity or behavioral information that public blockchain records do not provide.

## A.2 Robustness in the 28-Day Prospective Sample

Table A2 varies the candidate threshold, episode definition, outcome treatment, inference, and continuity restriction. All specifications use the 28-day prospective sample and retain the five-minute outcome as the main horizon.

Table A2: Robustness in the 28-day prospective sample

| Specification                         | Episodes | Markets | Coef. pp | Cluster SE |
|---------------------------------------|----------|---------|----------|------------|
| Baseline q95, 300-second window       | 173,240  | 2,271   | 1.610    | 0.150      |
| Candidate threshold q90               | 283,569  | 2,311   | 1.854    | 0.187      |
| Candidate threshold q99               | 51,588   | 2,017   | 1.180    | 0.176      |
| Episode window 60 seconds             | 273,686  | 2,288   | 1.198    | 0.139      |
| Episode window 900 seconds            | 121,091  | 2,241   | 1.621    | 0.172      |
| Large-trade episodes                  | 69,158   | 2,180   | 1.370    | 0.176      |
| Single-trade episodes                 | 112,746  | 2,252   | 0.735    | 0.185      |
| Outcome winsorized 1–99               | 173,240  | 2,271   | 1.399    | 0.128      |
| Outcome trimmed 1–99                  | 169,846  | 2,267   | 0.964    | 0.110      |
| Strict joint recorder segments        | 165,572  | 2,186   | 1.658    | 0.155      |
| Longest continuous segment            | 23,183   | 443     | 1.452    | 0.378      |
| Within one hour of a gap boundary     | 11,466   | 477     | 1.188    | 0.503      |
| Not within one hour of a gap boundary | 161,774  | 2,177   | 1.636    | 0.154      |

Notes: Every row reports the five-minute coefficient on log one plus top-5 depth pressure in percentage points. Models control for the pre-trade spread, log episode notional, market fixed effects, and UTC-hour fixed effects. Standard errors are clustered by market. The q95 and 300-second row is the baseline. The q90, q99, large-trade-only, single-trade, and 1–99 outcome rows vary the candidate or outcome definition. Episode-window rows change the maximum gap between consecutive candidate trades. Continuous-coverage rows require the complete event window to fall within one common trade-and-orderbook segment. The longest continuous segment covers 3.986 days. Boundary rows split episodes by whether the first trade lies within one hour of a coverage boundary. Episode-window and coverage rows are descriptive sensitivity analyses.

The q90 and q99 thresholds yield coefficients of 1.854 and 1.180 percentage points. Large-trade-only, single-trade, winsorized, and trimmed specifications also remain positive. Leave-one-market-out estimates range from 1.580 to 1.630 in the full sample and from 0.583 to 0.689 under isolation. Market-by-time and wallet clustering preserve the sign and precision. A 999-draw Webb bootstrap gives  $p = 0.001$  in both samples, while CV3 and approximate CR2 diagnostics give isolated-sample  $p$ -values below 0.0001 and 0.0020. The 60- and 900-second episode-window estimates are 1.198 and 1.621. Continuous-coverage and boundary estimates range from 1.188 to 1.658, showing magnitude sensitivity without a sign reversal.

### A.3 Prospective 24-Hour Depth Robustness

The following tables evaluate episode composition, outliers, market concentration, overlapping same-token episodes, candidate thresholds, episode windows, and orderbook freshness in the 24-hour prospective sample.

Table A3: Prospective depth-pressure robustness checks

| Horizon    | Single trade |         | Winsorized |         | Trimmed  |         | Leave-one-out range |
|------------|--------------|---------|------------|---------|----------|---------|---------------------|
| 60 seconds | 0.563***     | (0.217) | 0.292***   | (0.025) | 0.141*** | (0.031) | 0.388 to 0.482      |
| 5 minutes  | 0.587***     | (0.185) | 0.304***   | (0.023) | 0.218*** | (0.068) | 0.373 to 0.431      |
| 15 minutes | 0.525        | (0.320) | 0.326***   | (0.095) | 0.198*** | (0.044) | 0.236 to 0.461      |
| 1 hour     | 0.767***     | (0.277) | 0.565***   | (0.086) | 0.303**  | (0.128) | 0.600 to 0.735      |

Notes: Single trade restricts the sample to one-trade episodes. Winsorized and trimmed specifications treat the outcome and continuous regressors at the first and 99th percentiles. Leave-one-out reports the range of market and UTC-hour fixed-effects coefficients obtained by excluding one market at a time. Standard errors in the first three columns are clustered by market. Post-event horizons begin after the final episode trade. \*\*\*, \*\*, and \* denote significance at the 1, 5, and 10 percent levels.

At five minutes, the coefficient is 0.587 percentage points for single-trade episodes, 0.304 after winsorization, and 0.218 after trimming. The leave-one-market-out range is 0.373 to 0.431. Extreme observations contribute to the baseline magnitude, but no single market determines the coefficient.

Table A4: Event isolation and restricted wild-cluster inference

| Horizon    | Full pp          | Wild $p$ | Isolated pp     | Wild $p$ | Isolated N | Dollars |
|------------|------------------|----------|-----------------|----------|------------|---------|
| 60 seconds | 0.453*** (0.065) | 0.001    | 0.167** (0.044) | 0.041    | 1,690      | 146     |
| 5 minutes  | 0.410*** (0.046) | < 0.001  | 0.244** (0.138) | 0.042    | 956        | 214     |
| 15 minutes | 0.384 (0.127)    | 0.131    | 0.290 (0.373)   | 0.509    | 510        | 254     |
| 1 hour     | 0.674* (0.099)   | 0.052    | 0.365** (0.095) | 0.041    | 256        | 320     |

Notes: Full and isolated columns report market and UTC-hour fixed-effects coefficients on log one plus top-5 depth pressure. Market-clustered standard errors appear in parentheses. An episode is isolated if no other episode in the same token has a trading interval that intersects the window from the focal episode's first trade through the stated post-event horizon. Wild  $p$  reports a restricted wild-cluster bootstrap-t test with Rademacher weights, market clusters, 9,999 replications, and the null imposed. Dollars report the isolated implied value change per 100,000 dollars of exposure when depth pressure rises from 0.25 to 2. Stars use the wild-cluster  $p$ -value.

Isolation retains 1,690 observations at 60 seconds, 956 after five minutes, 510 after 15 minutes, and 256 after one hour. The isolated five-minute coefficient is 0.244 percentage points with a wild-cluster  $p$ -value of 0.042, equivalent to 214 dollars per 100,000 dollars of exposure. The 40 percent reduction from the full-sample coefficient indicates that repeated whale activity contributes to the baseline magnitude. The isolated sample shrinks sharply at longer horizons. The 15-minute estimate is imprecise, while the one-hour coefficient is statistically distinguishable from zero but is based on only 256 observations and no episode above twice displayed depth. These columns are therefore less informative about persistence than the five-minute result.

Table A5: Orderbook robustness grid for top-5 depth pressure

| Horizon    | Specs | Positive | Significant | Median pp | 10–90 pp       |
|------------|-------|----------|-------------|-----------|----------------|
| 60 seconds | 27    | 1.00     | 0.67        | 0.314     | 0.114 to 0.453 |
| 5 minutes  | 27    | 1.00     | 0.63        | 0.297     | 0.034 to 0.423 |
| 15 minutes | 27    | 1.00     | 0.44        | 0.262     | 0.144 to 0.393 |
| 1 hour     | 27    | 1.00     | 0.63        | 0.505     | 0.331 to 0.670 |

The grid varies the candidate threshold across q90, q95, and q99, the episode window across 60, 300, and 900 seconds, and maximum pre-trade snapshot age across 15, 30, and 60 seconds. All 27 coefficients are positive at every horizon. The median coefficient is 0.314 percentage points at 60 seconds, 0.297 after five minutes, 0.262 after 15 minutes, and 0.505 after one hour. The five-minute 10th-to-90th-percentile range is 0.034 to 0.423 percentage points.

## A.4 Short-Horizon Profitability

Profitability starts from the volume-weighted episode price, not the pre-episode midpoint. Midpoint P&L marks the episode to the future midpoint, best-quote P&L closes Buy episodes at the future best bid and Sell episodes at the future best ask, and full-size P&L requires displayed top-of-book quantity to cover the episode. Values are gross of platform and gas fees.

Table A6: Large-trade profitability across horizons

| Horizon    | Episodes | Midpoint P&L | Best-quote P&L | Full-size P&L | Positive (%) | Adjusted difference  |
|------------|----------|--------------|----------------|---------------|--------------|----------------------|
| 60 seconds | 1,469    | -65.5        | -142.9         | -107.6        | 8.8          | -29.6 [-75.6, 16.4]  |
| 5 minutes  | 1,466    | -98.0        | -170.3         | -122.4        | 13.6         | -10.7 [-94.8, 73.3]  |
| 15 minutes | 1,457    | -147.1       | -206.1         | -124.2        | 18.1         | -48.1 [-185.4, 89.1] |
| 1 hour     | 1,400    | -128.0       | -141.7         | -84.6         | 26.0         | 27.0 [-179.6, 233.6] |

Notes: P&L columns report average dollars per 100,000 dollars of binary payout exposure for episodes containing a market-specific 95th-percentile trade. Best-quote P&L closes at the future best bid or ask. Full-size P&L retains only episodes for which displayed quantity at that quote covers the episode size. Positive is the share of best-quote P&L observations above zero. Adjusted difference is the large-trade coefficient relative to ordinary trades, followed by a market-clustered 95 percent interval. Regressions include token and UTC-hour fixed effects and controls for trade direction, entry price, and squared entry price.

Average profit is negative at every horizon and under each marking rule. At five minutes, the large-trade loss is 98 dollars at the midpoint, 170 dollars at the best quote, and 122 dollars in the full-size subset. The preferred adjusted difference from ordinary trades is minus 11 dollars, with an interval from minus 95 to plus 73 dollars. Adding log trade size can produce positive sensitivity coefficients, but the large-trade level remains negative and the estimates are not stable across horizons, marking rules, clustering dimensions, or the preferred specification.

## A.5 Resolved-Market Profitability Robustness

Table A7: Resolved-market profitability across whale definitions

| Definition      | Role  | OOS difference | OOS 95% interval | Combined difference | Combined 95% interval |
|-----------------|-------|----------------|------------------|---------------------|-----------------------|
| Relative q90    | All   | 984            | [-479, 2,447]    | 1,006               | [124, 1,887]          |
|                 | Maker | -1,250         | [-3,498, 998]    | -304                | [-1,696, 1,088]       |
|                 | Taker | 2,432          | [293, 4,571]     | 1,682               | [396, 2,968]          |
| Relative q95    | All   | 1,043          | [-189, 2,274]    | 1,075               | [317, 1,832]          |
|                 | Maker | -1,025         | [-2,960, 909]    | -107                | [-1,322, 1,108]       |
|                 | Taker | 2,642          | [525, 4,760]     | 1,829               | [546, 3,111]          |
| Relative q97.5  | All   | 920            | [-367, 2,207]    | 989                 | [214, 1,764]          |
|                 | Maker | -1,662         | [-4,155, 831]    | -561                | [-2,094, 972]         |
|                 | Taker | 3,066          | [971, 5,161]     | 2,090               | [805, 3,375]          |
| Relative q99    | All   | 1,538          | [160, 2,915]     | 1,403               | [578, 2,229]          |
|                 | Maker | -359           | [-2,200, 1,481]  | 248                 | [-890, 1,385]         |
|                 | Taker | 3,454          | [902, 6,007]     | 2,423               | [860, 3,986]          |
| USD 500/2,500   | All   | 797            | [-656, 2,250]    | 841                 | [-32, 1,714]          |
|                 | Maker | -1,118         | [-3,286, 1,049]  | -242                | [-1,578, 1,094]       |
|                 | Taker | 2,303          | [280, 4,327]     | 1,543               | [323, 2,763]          |
| USD 1,000/5,000 | All   | 1,059          | [-218, 2,336]    | 997                 | [214, 1,780]          |
|                 | Maker | -772           | [-2,731, 1,188]  | -87                 | [-1,298, 1,125]       |
|                 | Taker | 2,563          | [436, 4,690]     | 1,705               | [414, 2,996]          |

Notes: Each difference is whale P&L minus nonwhale P&L in dollars per 100,000 dollars of payout exposure. Differences are averaged equally across markets. OOS denotes the 30 markets selected before their fill outcomes were analyzed. Combined adds the 20 highest-volume eligible markets. Intervals use Student- $t$  critical values across markets.

Relative thresholds vary from q90 to q99, and the two absolute definitions use fill and wallet-market cutoffs. Every taker estimate is positive in both samples, all twelve taker intervals exclude zero, and combined-sample taker differences range from 1,543 to 2,423 dollars. Maker estimates remain unstable. Leave-one-market-out estimates preserve a positive taker difference under every definition.

Table A8: Resolved-market profitability by pre-specified volume stratum

| Role  | Volume-rank stratum | Markets | Difference | 95% interval     | Positive markets |
|-------|---------------------|---------|------------|------------------|------------------|
| All   | 21–50               | 8       | 1,870      | [515, 3,224]     | 7                |
|       | 51–100              | 8       | 768        | [-1,945, 3,480]  | 7                |
|       | 101–175             | 7       | 500        | [-4,664, 5,663]  | 5                |
|       | 176–250             | 7       | 955        | [-1,044, 2,955]  | 5                |
| Maker | 21–50               | 8       | 114        | [-2,817, 3,045]  | 6                |
|       | 51–100              | 8       | 708        | [-1,516, 2,932]  | 6                |
|       | 101–175             | 7       | -6,690     | [-13,238, -142]  | 2                |
|       | 176–250             | 7       | 1,357      | [-1,126, 3,840]  | 5                |
| Taker | 21–50               | 8       | 3,067      | [1,028, 5,107]   | 8                |
|       | 51–100              | 8       | 313        | [-2,341, 2,966]  | 6                |
|       | 101–175             | 7       | 6,871      | [-1,643, 15,384] | 6                |
|       | 176–250             | 7       | 592        | [-3,414, 4,598]  | 5                |

Notes: The table uses the preferred within-market q95 whale definition in the 30-market out-of-sample sample. Differences are whale P&L minus nonwhale P&L in dollars per 100,000 dollars of payout exposure. Student- $t$  intervals treat markets as independent observations. Volume ranks refer to the eligible 250-market universe.

The taker difference is positive in every pre-specified volume stratum and equals 3,067 dollars among ranks 21–50. Other cells contain seven or eight markets and are less precise. A continuous volume diagnostic has an HC1 interval from minus 7,909 to 7,265 dollars and provides no evidence of a monotonic volume gradient.

## A.6 Ex-Ante Threshold and Timing Robustness

Table A9: Ex-ante profitability across relative thresholds

| Design             | Threshold | OOS taker diff.      | OOS interaction      | Combined taker diff. | Combined interaction |
|--------------------|-----------|----------------------|----------------------|----------------------|----------------------|
| Initial 20 percent | q90       | 1,175 [-981, 3,331]  | 3,016 [-754, 6,786]  | 756 [-527, 2,039]    | 1,531 [-778, 3,839]  |
| Initial 20 percent | q95       | 1,661 [-527, 3,848]  | 3,929 [87, 7,771]    | 1,044 [-267, 2,356]  | 2,136 [-223, 4,495]  |
| Initial 20 percent | q99       | 3,035 [571, 5,499]   | 5,088 [1,041, 9,134] | 1,904 [380, 3,428]   | 2,947 [433, 5,460]   |
| Expanding          | q90       | 1,585 [-699, 3,869]  | 3,860 [271, 7,449]   | 1,016 [-344, 2,376]  | 1,974 [-246, 4,194]  |
| Expanding          | q95       | 2,358 [292, 4,423]   | 4,276 [710, 7,843]   | 1,493 [234, 2,751]   | 2,373 [167, 4,579]   |
| Expanding          | q99       | 3,139 [1,023, 5,255] | 4,312 [623, 8,002]   | 1,938 [597, 3,279]   | 2,418 [133, 4,704]   |

Notes: Each cell reports a market-equal difference followed by its 95 percent Student-*t* interval. Taker difference compares large and small aggressive orders. Interaction subtracts the corresponding maker difference. Initial-period thresholds use the first 20 percent of distinct trading timestamps. Expanding thresholds use only earlier timestamp blocks. All magnitudes are dollars per 100,000 dollars of payout exposure.

Under the expanding design, the combined taker difference rises from 1,016 dollars at q90 to 1,938 dollars at q99, and the interaction rises from 1,974 to 2,418 dollars. The q95 and q99 intervals exclude zero for both estimates. The initial-period design is less precise at q90 and q95, while its q99 estimates remain positive and statistically distinguishable from zero.

Table A10: Ex-ante fill-size profitability across market-time regions

| Sample      | Market phase        | Taker diff. | 95% interval    | Wild $p$ | Interaction | 95% interval     | Wild $p$ |
|-------------|---------------------|-------------|-----------------|----------|-------------|------------------|----------|
| OOS 30      | Early, decile 2     | 7,728       | [3,104, 12,351] | 0.001    | 12,739      | [3,225, 22,254]  | 0.009    |
|             | Middle, deciles 3–8 | 578         | [-2,552, 3,709] | 0.721    | 2,838       | [213, 5,463]     | 0.035    |
|             | Late, deciles 9–10  | 1,861       | [352, 3,370]    | 0.012    | 1,784       | [-541, 4,108]    | 0.139    |
| Combined 50 | Early, decile 2     | 3,378       | [-439, 7,195]   | 0.081    | 3,737       | [-3,992, 11,465] | 0.343    |
|             | Middle, deciles 3–8 | 369         | [-1,511, 2,248] | 0.704    | 1,485       | [-192, 3,162]    | 0.077    |
|             | Late, deciles 9–10  | 1,505       | [560, 2,451]    | 0.001    | 1,519       | [-22, 3,060]     | 0.052    |

Notes: Market phases are constructed from distinct trading-timestamp deciles. Decile one is used only for initial calibration. Each later decile uses thresholds based exclusively on earlier deciles. Decile-level differences are weighted equally within each market and phase. Wild  $p$ -values use 9,999 market-level Rademacher sign flips. All magnitudes are dollars per 100,000 dollars of payout exposure.

The out-of-sample taker difference is 7,728 dollars in the early phase and 1,861 dollars in the late phase. The combined late-phase difference is 1,505 dollars, with an interval from 560 to 2,451 dollars. Middle-phase direct taker differences are small and imprecise. The aggregate result is not confined to settlement, but its strength varies over the market life cycle.

## A.7 Out-of-Sample Absolute-Return Concentration

Table A11: Out-of-sample large-taker profitability diagnostics

| Specification            | Net P&L | 95% bootstrap interval | LOO range      | Diff. | Wild $p$ |
|--------------------------|---------|------------------------|----------------|-------|----------|
| q90, all fills           | 1,446   | [-19, 2,994]           | [992, 1,659]   | 1,585 | 0.181    |
| q95, all fills           | 1,835   | [249, 3,520]           | [1,351, 2,083] | 2,358 | 0.018    |
| q99, all fills           | 2,989   | [884, 5,086]           | [2,414, 3,360] | 3,139 | 0.000    |
| q95, price 5–95 cents    | 3,224   | [403, 5,851]           | [2,488, 3,731] | 4,222 | 0.001    |
| q95, price 10–90 cents   | 4,008   | [816, 6,912]           | [3,261, 4,637] | 3,709 | 0.067    |
| q95, more than 24 hours  | 412     | [-320, 1,237]          | [209, 559]     | 2,354 | 0.032    |
| q95, excluding sports    | 262     | [-234, 809]            | [77, 375]      | 1,162 | 0.030    |
| q95, excluding top three | 599     | [-548, 1,949]          | [217, 863]     | 1,925 | 0.016    |

Notes: Net P&L is the exposure-weighted settlement return of large taker orders under the stated threshold. Recorded exchange fees are included. Bootstrap intervals resample whole markets with 9,999 draws. LOO reports the range after excluding each market in turn. Diff. is the market-equal large-minus-other taker difference. Wild  $p$ -values use 9,999 market-level Rademacher sign flips. Magnitudes are dollars per 100,000 dollars of binary payout exposure. Price and subgroup analyses are diagnostic and were not preregistered.

The q95 net level is 1,835 dollars, with a 95 percent market-cluster bootstrap interval from 249 to 3,520 dollars. Seven sports markets yield 5,450 dollars, while the estimate outside sports is 262 dollars with an interval from minus 234 to 809 dollars. Excluding the three largest positive contributors reduces the level to 599 dollars, and trades more than seven days before close earn minus 62 dollars. Relative order-size performance is broader than the positive absolute level, which is concentrated in event-proximate markets.

## A.8 Cross-Platform Threshold and Family Robustness

Table A12: Cross-platform threshold and family robustness

| <i>Panel A: Alternative trade-size thresholds</i> |                |                   |                       |          |
|---------------------------------------------------|----------------|-------------------|-----------------------|----------|
| Threshold                                         | Settlement P&L | Kalshi 30 minutes | Polymarket 30 minutes |          |
| q90                                               | 337 (0.724)    | 530 (0.013)       | 514 (0.006)           |          |
| q95                                               | 365 (0.708)    | 300 (0.172)       | 448 (0.031)           |          |
| q99                                               | 338 (0.762)    | 197 (0.200)       | 480 (0.012)           |          |
| <i>Panel B: Leave-one-family-out q95 ranges</i>   |                |                   |                       |          |
| Leave-one-family-out outcome                      | Minimum        | Maximum           | Min. $p$              | Max. $p$ |
| Settlement P&L                                    | -126           | 708               | 0.491                 | 0.897    |
| Kalshi, 30 minutes                                | 140            | 423               | 0.127                 | 0.336    |
| Polymarket, 30 minutes                            | 284            | 645               | 0.021                 | 0.070    |

Notes: The upper panel reports differences in dollars per 100,000 dollars of binary exposure with event-level randomization  $p$ -values in parentheses. The lower panel reports the range across 15 specifications that exclude one validation family at a time. The price-distance rule retains observations only when the measured price is within ten minutes of the target horizon. Same-day congressional primaries within California, Georgia, and New York are collapsed into state-level macro-events.

Settlement P&L remains statistically indistinguishable from zero at q90, q95, and q99. The 30-minute Polymarket response is positive and statistically distinguishable from zero at all three thresholds. The q95 leave-one-family-out range is 284 to 645 dollars, with a maximum randomization  $p$ -value of 0.070. No single family changes the sign, although event composition affects precision.

## A.9 Cross-Platform Leads and Source-Price Controls

Table A13: Cross-platform leads and source-price controls

| Specification                              | Markets | Events | Estimate | 95% interval  | Wild $p$ / Holm $p$ |
|--------------------------------------------|---------|--------|----------|---------------|---------------------|
| Prior 30-minute Polymarket move            | 246     | 33     | 262      | [-362, 887]   | 0.673 / –           |
| Lead-available raw response                | 246     | 33     | 448      | [17, 878]     | 0.033 / 0.098       |
| + Polymarket pretrend                      | 246     | 33     | 374      | [-31, 778]    | 0.047 / 0.140       |
| Full-price sample, raw                     | 109     | 27     | 431      | [-182, 1,044] | 0.142 / 0.403       |
| + Polymarket and Kalshi pretrends          | 109     | 27     | 517      | [-160, 1,194] | 0.093 / 0.280       |
| + Same-horizon Kalshi return               | 109     | 27     | 130      | [-222, 482]   | 0.412 / 0.823       |
| Full controls across trade-size thresholds |         |        |          |               |                     |
| q90                                        | 110     | 27     | 155      | [-254, 564]   | 0.421 / 1.000       |
| q95                                        | 109     | 27     | 130      | [-222, 482]   | 0.412 / 0.823       |
| q99                                        | 102     | 25     | 176      | [-170, 522]   | 0.311 / 0.933       |

Notes: Estimates and intervals are dollars per 100,000 dollars of binary exposure. Wild  $p$ -values use 9,999 Rademacher sign flips at the macro-event level. Holm  $p$ -values adjust across the three post-trade horizons within a specification and threshold. The prior-movement row is a lead placebo. The same-horizon Kalshi return is a post-trade transmission variable and not an exogenous control. Price matches must lie within ten minutes of their target timestamps.

Prior Polymarket movement accounts for 74 dollars of the 448 dollar q95 response, while the lead itself is insignificant. In the common source-price sample, the same-horizon Kalshi return accounts for most of the conditional response. Across 14 leave-one-family-out specifications, the fully controlled estimate ranges from minus 121 to 258 dollars and every event-level randomization  $p$ -value exceeds 0.230.

## A.10 Equivalence and Detectable Settlement Returns

Table A14: Settlement-return equivalence and minimum detectable effects

| Sample                                | Estimate | 90% interval    | MDE 80% | $\pm 1,000$ | $\pm 2,000$ |
|---------------------------------------|----------|-----------------|---------|-------------|-------------|
| Kalshi q95 large-minus-other          | 365      | [-1,167, 1,897] | 2,610   | No          | Yes         |
| Kalshi q95 large absolute             | 1,091    | [-2,183, 4,365] | 5,577   | No          | No          |
| Polymarket combined large-minus-other | 1,493    | [463, 2,523]    | 1,754   | No          | No          |
| Polymarket combined large absolute    | -48      | [-539, 564]     | 949     | Yes         | Yes         |
| Polymarket OOS large absolute         | 1,835    | [478, 3,257]    | 2,364   | No          | No          |

Notes: Estimates, intervals, margins, and minimum detectable effects are dollars per 100,000 dollars of payout exposure. Equivalence is established when both one-sided tests reject at five percent. Kalshi estimates average contract differences within conservative macro-event clusters. Polymarket large-minus-other estimates are market-equal. Absolute Polymarket intervals use 9,999 market-cluster bootstrap draws. Returns are gross of unrecorded trading costs.

The q95 Kalshi difference is equivalent to zero within a 2,000 dollar margin, but not within a 1,000 dollar margin. Its 2,610 dollar minimum detectable effect leaves moderate relative alpha unresolved. The combined Polymarket large-taker level is equivalent to zero within 1,000 dollars and has a 949 dollar minimum detectable effect, while the positive held-out level shows cohort heterogeneity.

## A.11 Matching Specification Grid

Table A15: Range of whale-control differences across 16 matching specifications

| Horizon    | Minimum difference | Maximum difference |
|------------|--------------------|--------------------|
| 60 seconds | 0.0009             | 0.0010             |
| 5 minutes  | 0.0010             | 0.0012             |
| 15 minutes | 0.0010             | 0.0012             |
| 1 hour     | 0.0007             | 0.0012             |

The grid varies the number of controls across 1, 3, 5, and 10 and the exclusion window across 30 minutes, one hour, two hours, and six hours. The whale-control difference remains positive from 60 seconds to one hour in all 16 specifications. This stability addresses control-count and exclusion-window choices, not differences in contemporaneous information arrival.

## A.12 Exploratory Temporal Stability

Figure A1 displays calendar partitions, rolling seven-day windows, and expanding windows within the later, market-disjoint outcome sample. The windows overlap and were constructed after the outcomes were observed. They therefore describe temporal breadth but do not provide independent replications or additional confirmatory tests.

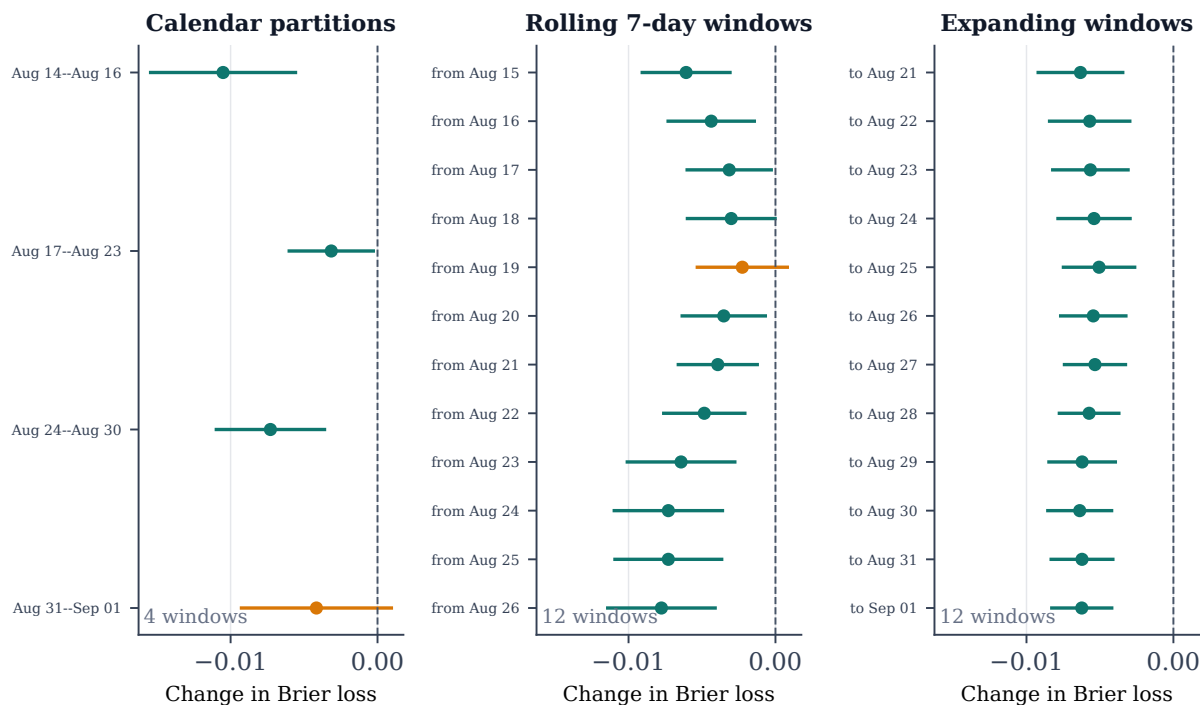


Figure A1: Brier-loss changes across alternative time windows

*Notes:* Points are market-equal post-minus-pre Brier-loss changes. Bars are normal 95 percent intervals based on event-family-clustered standard errors. Negative values indicate lower realized squared forecast loss after the episode. The analysis is descriptive and was constructed after outcomes were observed. Calendar, rolling, and expanding windows overlap.

## A.13 Additional Descriptive Figures

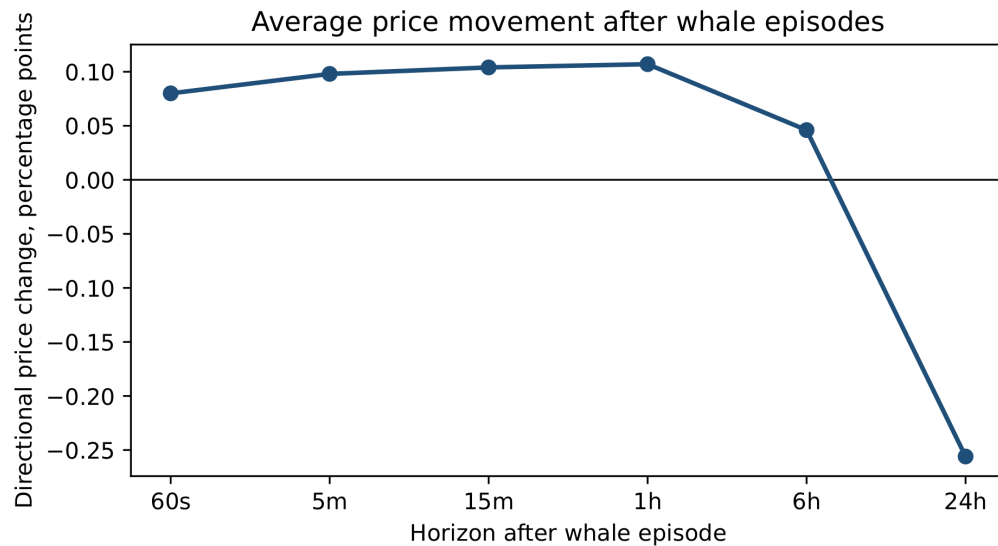


Figure A2: Average directional price movement in the retrospective sample

*Notes:* Positive values indicate movement in the trade direction. Sample size falls from 440 episodes at short horizons to 255 episodes after 24 hours. The changing support limits long-horizon comparisons.

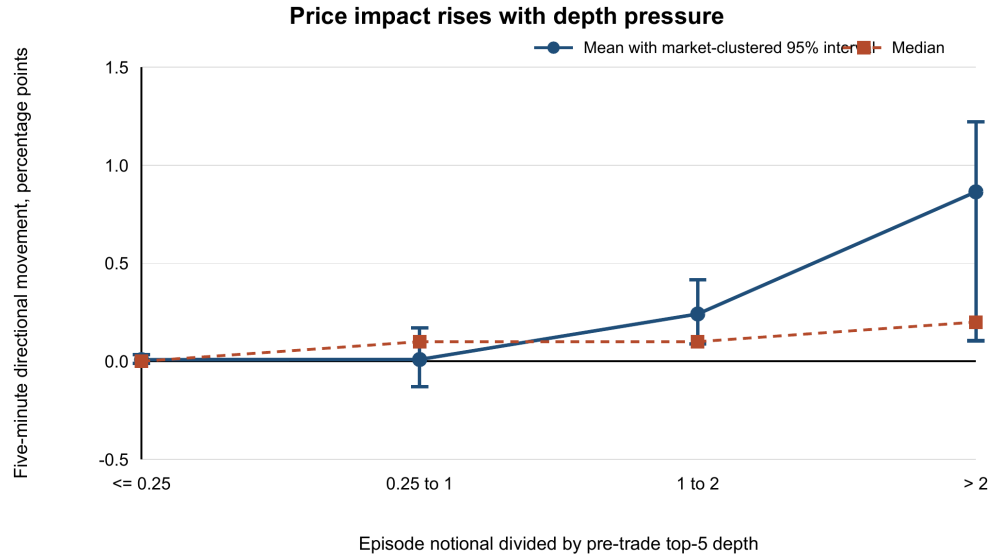


Figure A3: Five-minute price response by depth pressure in the prospective 24-hour sample

*Notes:* The figure reports bin means with market-clustered 95 percent bootstrap intervals and bin medians. Depth pressure is total episode notional divided by pre-trade top-5 displayed depth on the relevant book side.

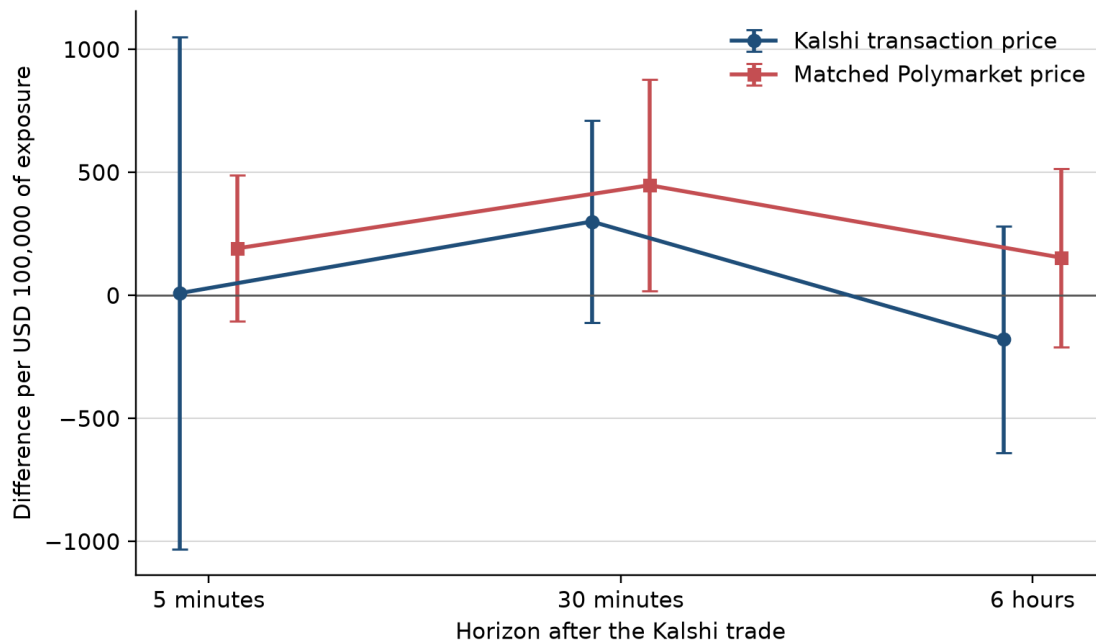


Figure A4: Large-minus-other trade differences after Kalshi transactions

Notes: Points report q95 event-equal estimates in dollars per 100,000 dollars of exposure. Bars show 95 percent Student- $t$  intervals.

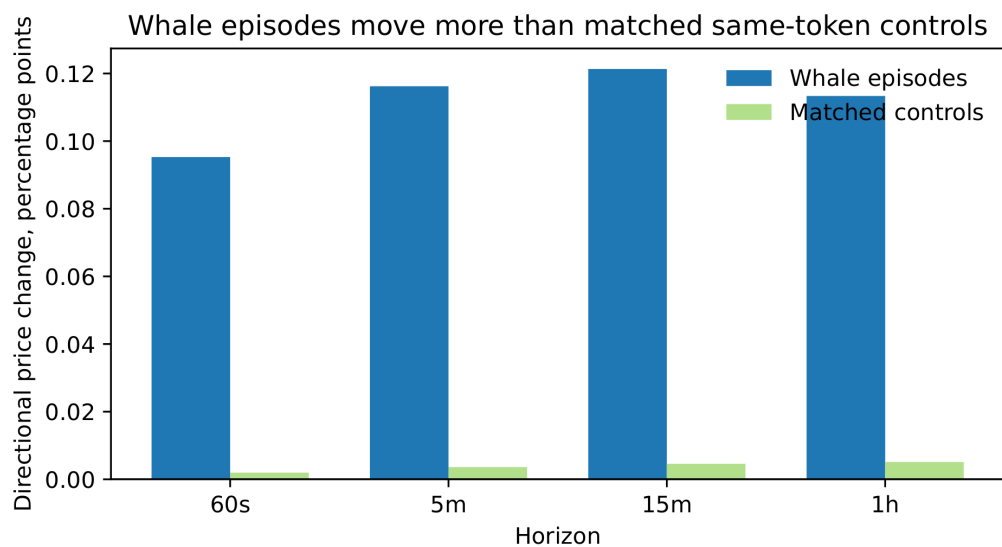


Figure A5: Large-order episodes versus matched same-token timestamps

Notes: Control timestamps are selected within the same token, matched on the base price, and placed away from large-order windows. Reused timestamps create dependence and the comparison is descriptive.

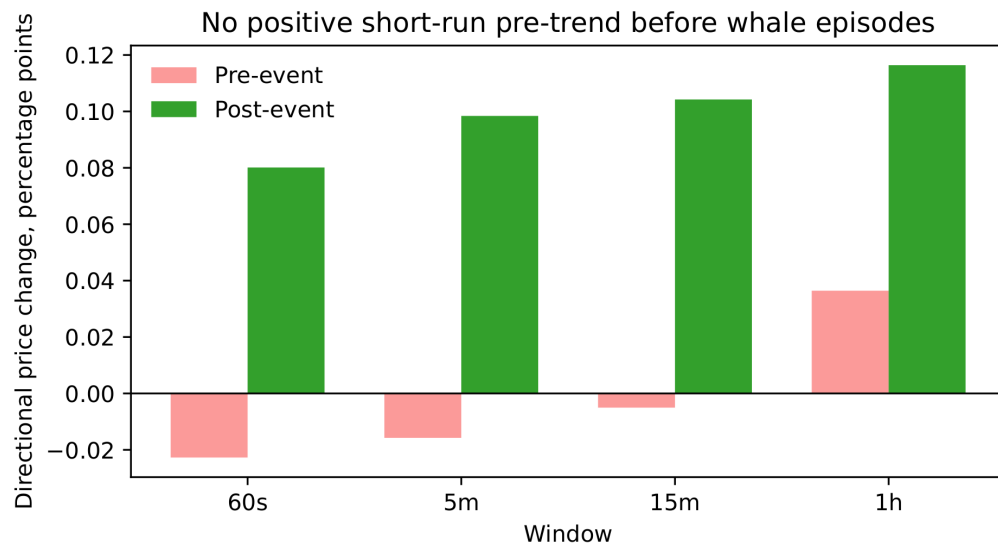


Figure A6: Pre-event and post-event directional movement

*Notes:* Short-run movement appears after large-order episodes and not before them. The comparison does not distinguish mechanical pressure from information arriving at the episode timestamp.

## References

- Berg, J., and T. Rietz. 2006. The Iowa Electronic Markets: Stylized Facts and Open Issues. In R. W. Hahn and P. C. Tetlock, eds., *Information Markets: A New Way of Making Decisions*, 142–169. Washington, DC: AEI Press.
- Berg, J., F. Nelson, and T. Rietz. 2008. Prediction Market Accuracy in the Long Run. *International Journal of Forecasting* 24(2): 285–300. doi:10.1016/j.ijforecast.2008.03.007.
- Biais, B., P. Hillion, and C. Spatt. 1995. An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse. *Journal of Finance* 50(5): 1655–1689. doi:10.1111/j.1540-6261.1995.tb05192.x.
- Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78(1): 1–3.
- Chakraborty, M., and S. Das. 2014. On Manipulation in Prediction Markets When Participants Influence Outcomes Directly. arXiv:1407.7015. <https://arxiv.org/abs/1407.7015>.
- Dai, D., R. Jia, and S. Yu. 2026. Settlement Manipulation in Prediction Markets. arXiv:2606.31675. doi:10.48550/arXiv.2606.31675. <https://arxiv.org/abs/2606.31675>.
- Erikson, R., and C. Wlezien. 2008. Are Political Markets Really Superior to Polls as Election Predictors? *Public Opinion Quarterly* 72(2): 190–215. doi:10.1093/poq/nfn010.
- Erikson, R., and C. Wlezien. 2012. Markets vs. Polls as Election Predictors: An Historical Assessment. *Electoral Studies* 31(3): 532–539. doi:10.1016/j.electstud.2012.04.008.
- Forsythe, R., F. Nelson, G. Neumann, and J. Wright. 1992. Anatomy of an Experimental Political Stock Market. *American Economic Review* 82(5): 1142–1161.
- Forsythe, R., T. Rietz, and T. Ross. 1999. Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets. *Journal of Economic Behavior and Organization* 39(1): 83–110. doi:10.1016/S0167-2681(99)00027-X.
- Gjerstad, S. 2005. Risk Aversion, Beliefs, and Prediction Market Equilibrium. University of Arizona Department of Economics Working Paper 04-17.
- Glosten, L., and P. Milgrom. 1985. Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics* 14(1): 71–100. doi:10.1016/0304-405X(85)90044-3.
- Graefe, A., and J. Armstrong. 2011. Comparing Face-to-Face Meetings, Nominal Groups, Delphi and Prediction Markets on an Estimation Task. *International Journal of Forecasting* 27(1): 183–195. doi:10.1016/j.ijforecast.2010.05.004.

- Hanson, R. 2003. Combinatorial Information Market Design. *Information Systems Frontiers* 5(1): 107–119. doi:10.1023/A:1022058209073.
- Hanson, R. 2007. Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation. *Journal of Prediction Markets* 1(1): 3–15. doi:10.5750/jpm.v1i1.417.
- Hanson, R., R. Oprea, and D. Porter. 2006. Information Aggregation and Manipulation in an Experimental Market. *Journal of Economic Behavior and Organization* 60(4): 449–459. doi:10.1016/j.jebo.2004.09.011.
- Hasbrouck, J. 1991. Measuring the Information Content of Stock Trades. *Journal of Finance* 46(1): 179–207. doi:10.1111/j.1540-6261.1991.tb03749.x.
- Kyle, A. 1985. Continuous Auctions and Insider Trading. *Econometrica* 53(6): 1315–1335. doi:10.2307/1913210.
- Le, N. A. 2026. Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets. arXiv:2602.19520. doi:10.48550/arXiv.2602.19520. <https://arxiv.org/abs/2602.19520>.
- Manski, C. 2006. Interpreting the Predictions of Prediction Markets. *Economics Letters* 91(3): 425–429. doi:10.1016/j.econlet.2006.01.004.
- Murphy, A. H. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* 12(4): 595–600.
- Nechepurenko, M. 2026a. Fill-Side Behavioral Concentration on Polymarket: Identification Limits under Record-Level Attribution. arXiv:2605.11640. doi:10.48550/arXiv.2605.11640. <https://arxiv.org/abs/2605.11640>.
- Nechepurenko, M. 2026b. Manipulation, Informed Trading, and Regulation in Leveraged Event-Linked Markets. arXiv:2605.10486. doi:10.48550/arXiv.2605.10486. <https://arxiv.org/abs/2605.10486>.
- Page, L., and R. Clemen. 2013. Do Prediction Markets Produce Well-Calibrated Probability Forecasts? *The Economic Journal* 123(568): 491–513. doi:10.1111/j.1468-0297.2012.02561.x.
- Polgreen, P. M., F. D. Nelson, and G. R. Neumann. 2007. Use of Prediction Markets to Forecast Infectious Disease Activity. *Clinical Infectious Diseases* 44(2): 272–279. doi:10.1086/510427.
- Qin, B., and R. Yang. 2026. Polymarket-v1 Database. arXiv:2606.04217. doi:10.48550/arXiv.2606.04217. <https://arxiv.org/abs/2606.04217>.
- Rhode, P., and K. Strumpf. 2004. Historical Presidential Betting Markets. *Journal of Economic Perspectives* 18(2): 127–141. doi:10.1257/0895330041371277.

- Rhode, P., and K. Strumpf. 2008. Historical Political Futures Markets: An International Perspective. NBER Working Paper No. 14377. doi:10.3386/w14377. <https://www.nber.org/papers/w14377>.
- Rosenbloom, E., and W. Notz. 2006. Statistical Tests of Real-Money versus Play-Money Prediction Markets. *Electronic Markets* 16(1): 63–69. doi:10.1080/10196780500491303.
- Servan-Schreiber, E., J. Wolfers, D. Pennock, and B. Galebach. 2004. Prediction Markets: Does Money Matter? *Electronic Markets* 14(3): 243–251. doi:10.1080/1019678042000245254.
- Shen, Y., Y. Jin, S. Wu, Y. Wang, and J. Chen. 2026. The Ghosts of Polymarket: When Off-Chain Matches Meet On-Chain Reverts. arXiv:2606.16852. doi:10.48550/arXiv.2606.16852. <https://arxiv.org/abs/2606.16852>.
- Spann, M., and B. Skiera. 2003. Internet-Based Virtual Stock Markets for Business Forecasting. *Management Science* 49(10): 1310–1326. doi:10.1287/mnsc.49.10.1310.17314.
- Wolfers, J., and E. Zitzewitz. 2004. Prediction Markets. *Journal of Economic Perspectives* 18(2): 107–126. doi:10.1257/0895330041371321.
- Wolfers, J., and E. Zitzewitz. 2006. Interpreting Prediction Market Prices as Probabilities. NBER Working Paper No. 12200. doi:10.3386/w12200. <https://www.nber.org/papers/w12200>.