Kalshi Research

Who Calibrates Markets? Evidence from Trader Micro-Data

Nicole Kagan and Rubens Baiocchi

Working Paper

October 2026

Abstract

Past academic work has shown prediction markets, and Kalshi markets in particular, are well-calibrated. What is not well understood, however, is why. In this paper, we use evidence from more than 1 billion trades executed from October 2025 to September 2026 on the world's largest prediction market to suggest that there may exist, in line with the "superforecaster" literature, a small group of identifiably highly skilled traders who meaningfully contribute to price discovery. Using a Brier skill score and analysing only takers (excluding market-makers) illustrates that under a strict false discovery threshold, 0.15 percent of traders forecast significantly better than the markets they trade against. This skill persists out of sample and is largely orthogonal to profit. After a skilled taker's trade, the weighted midpoint moves 1.7 cents toward their direction by the next order and 3.3 cents by close, which constitutes most of the 3.8-cent gap between the pre-trade midpoint and the outcome. A difference-in-differences design against volume-matched markets finds the same, with each additional skilled trader having diminishing returns. Markets with more skilled traders tend to also be more accurate on average. Overall, prices move toward the outcome after the trades of skilled traders, not necessarily profitable ones. Skilled trading thus joins volume, the total number of traders, and time to resolution as a source of accuracy and calibration in prediction markets.

Keywords: prediction markets, price discovery, forecast skill, Brier score, market microstructure.
JEL Classification: C12, G14, D83.

1 Introduction

Studies of prediction markets have shown them to be accurate and reasonably well calibrated (Wolfers and Zitzewitz, 2004, 2006; Berg, Nelson, and Rietz, 2008; Snowberg, Wolfers, and Zitzewitz, 2013; Reichenbach and Walther, 2026; Diercks, Katz, and Wright, 2026). On Kalshi, we showed that calibration improves with trading volume, with the number of participating traders, and with time (Kagan and Baiocchi, 2026). However, volume and the total number of traders are only rough measures of the informational content of markets. Markets populated only by traders who buy and sell contracts at random have no resolution by construction. Their prices cannot separate events that happen from events that don't. Therefore, the quality of the order flow should be the more important indicator (Glosten and Milgrom, 1985). This is not to say that a large share of skilled traders is required to have accurate markets. On the contrary, Oliven and Rietz (2004) find that many traders on the Iowa Electronic Markets violated the law of one price, yet market probabilities remained largely accurate. A skilled minority may suffice.

This assertion rests on the market microstructure literature, which divides participants into three classes: skilled traders, noise traders, and market makers (Glosten and Milgrom, 1985; Kyle, 1985). In this framework, skilled traders make the directional forecasts and so impose adverse selection on market makers. On Kalshi, Bartlett and O'Hara (2026) show that market makers recoup these losses from noise traders. For this mechanism to work, prices cannot be perfectly efficient, since skilled traders need compensation for the private cost of acquiring their information (Grossman and Stiglitz, 1980). Their forecasts should therefore be more accurate than market prices. In the language of the forecasting literature, these traders are the market's superforecasters (Tetlock and Gardner, 2015).

Finding them empirically is a problem of separating skill from luck. In the classical finance literature on fund manager returns, a track record counts as evidence of skill only once it is measured against an appropriate benchmark and tested against its own sampling noise (Jensen, 1968). Fama and French (2010) and Barras, Scaillet, and Wermers (2010) show that, in a large enough cross-section, luck alone will always throw up impressive results. Contracts in prediction markets settle at zero or one, making outcomes frequently observable, yet that also sharpens these difficulties. A trader who buys only 97-cent favorites can go many events without a loss yet have no skill at all. Moreover, a platform with millions of accounts will produce thousands of apparent superforecasters by chance. In prediction markets, Gómez-Cram, Guo, Jensen et al. (2026) randomize the sign of each trader's position to measure skill as profitability.

Reichenbach and Walther (2026) suggest that consistently profitable traders exploit the biases of noise traders. One is the favorite–longshot bias (Bürgi, Deng, and Whelan, 2026; Page and Clemen, 2013). However, profitable traders need not have an accuracy edge and may earn their returns simply by providing liquidity (Gu, Kagan, Sun et al., 2026; Akey, Grégoire, Harvie et al., 2026). Della Vedova (2026) finds that forecasting accuracy and execution are nearly independent on Polymarket, and that execution rather than accuracy determines who profits. Relative to this literature, we measure skill directly as forecasting accuracy using the Brier skill score (BSS). We then measure how prices move after skilled traders' trades, following Huang and Stoll (1996) and Glosten and Harris (1988), and complement this with a difference-in-differences design (DiD) around skilled entry.

The rest of the paper proceeds as follows. Section 2 describes the data and how the market environment connects with the literature. Section 3 defines our measure of skill and separates skill from luck by comparing each trader's score with the scores the same trades would earn if market outcomes were drawn at random from prices, controlling the false discovery rate with the Benjamini–Hochberg (BH) procedure. Section 4 identifies the skilled traders, describes them, and decomposes the sources of their skill. Section 5 links them to market accuracy. Section 6 addresses the study's limitations. Section 7 concludes and discusses what the results imply for how participants supply calibration and how information reaches prices.

2 Data and Market Environment

2.1 Sample

Kalshi is a Designated Contract Market regulated by the Commodity Futures Trading Commission. Its core product is the event contract, which pays \$1 if a specified outcome occurs and \$0 otherwise. The contract trades between \$0.01 and \$0.99 in one-cent increments, and some markets also allow sub-cent increments below 10 cents and above 90 cents. The exchange matches each trade on a central limit order book between a maker, whose resting order supplies liquidity, and a taker, whose order crosses the spread and removes it.

Our sample is the universe of trades executed on Kalshi between October 1, 2025, and September 2, 2026, in all non-Exotic markets. We augment each trade with the top-of-book bid and ask quotes and their respective resting sizes prevailing immediately before it. We exclude trades executed against a one-sided book (2.20 percent of trades), trades in markets that had not settled by the end of the sample (0.84 percent), trades in markets that did not resolve to a binary outcome (0.21 percent), and block trades (0.11 percent). The analysis sample comprises 1,083,797,223 trades across 535,019 events and 2,353,157 markets. These trades were placed by 9,656,615 users, whom we refer to interchangeably as traders.

3 Measuring Skill

3.1 Prices as Forecasts

Since Glosten and Milgrom (1985) and Kyle (1985), the microstructure literature has treated market makers as a distinct class of participants, and empirical work on prediction markets has maintained the same distinction (Gómez-Cram, Guo, Jensen et al., 2026; Akey, Grégoire, Harvie et al., 2026; Bartlett and O'Hara, 2026). Because we aim to measure pure price-discovery and forecasting skill, we score only the taker side of each trade. This exclusion follows the standard view of the taker as the side with new information, whose trades impose adverse selection on market makers.

Under fairly general conditions, a market price can be read both as the traders' mean belief and admit a literal probabilistic interpretation (Wolfers and Zitzewitz, 2006). In the context of this study, we use the weighted midpoint between the best bid and ask as the market's forecast (Stoikov, 2018). By transacting at the own-side price p_i, the taker makes a competing forecast of the same event. That trade signals a view that the true probability is at least p_i and that

the market is not perfectly efficient (Grossman and Stiglitz, 1980). We therefore score p_i as a conservative estimate of the taker's belief. The market and taker forecasts both lie in $[0, 1]$ and settle on the same binary outcome $Y_i \in \{0, 1\}$, so they can be compared by any proper scoring rule (Gneiting and Raftery, 2007).

A skilled taker contributes to price discovery by thinning the order book and moving the midpoint itself. This skill is distinct from market making. It can arise either from speed, reacting to public information before the book does, or from a more accurate model of the world (Tetlock and Gardner, 2015). In both cases, skill means that the taker's belief sits closer to the truth than the market's, on average.

Defining skill in this manner is deliberately decoupled from the question of profit. A trader can earn a positive return by supplying liquidity or by sizing positions well. Akey, Grégoire, Harvie et al. (2026) and Della Vedova (2026), for example, find that liquidity provision is the strongest predictor of profitability on other exchanges. Conversely, as Appendix A.2 shows, a skilled taker who pushes the price all the way to the truth can score better on accuracy than one who stops short and earns more per contract.

3.2 The Brier Skill Score

We use the Brier score (Brier, 1950) in skill-score form (Glahn and Jorgensen, 1970; Wilks, 2011) as a measure of trader ability. All prices are stated on the trader's own side, so that $Y_i = 1$ when the market resolves in favor of the side taken by the taker. Each trade is scored independently, at its own execution price, against the market's weighted midpoint immediately before it. With best bid and ask prices P^{bid} and P^{ask} and resting quantities Q^{bid} and Q^{ask}, the quantity-weighted midpoint is:

$$m_w \;=\; \frac{P^{\text{ask}}\, Q^{\text{bid}} + P^{\text{bid}}\, Q^{\text{ask}}}{Q^{\text{bid}} + Q^{\text{ask}}} \tag{1}$$

The ask price is weighted by the resting quantity at the bid, and the bid price by the resting quantity at the ask, so the midpoint leans toward the side with less resting volume. Appendix A.4 explains why it is a better reference than the plain midpoint. Per trade, the trader's accuracy gain over the market is the midpoint's squared error minus the trader's.

$$S_i \;=\; (m_{w,i} - Y_i)^2 - (p_i - Y_i)^2 \tag{2}$$

Trades are weighted throughout by their contract size w_i. For trader j with set of taker trades $\mathcal{T}_j$, the trader-level score sums these gains and scales them by the market's squared error on the same trades:

$$\text{BSS}_j \;=\; 1 - \frac{\sum_{i \in \mathcal{T}_j} w_i\, (p_i - Y_i)^2}{\sum_{i \in \mathcal{T}_j} w_i\, (m_{w,i} - Y_i)^2} \;=\; \frac{\sum_{i \in \mathcal{T}_j} w_i\, S_i}{\sum_{i \in \mathcal{T}_j} w_i\, (m_{w,i} - Y_i)^2} \tag{3}$$

The score is a ratio of sums rather than a mean of per-trade ratios (Wilks, 2011). BSS_j equals zero when the trader's prices are exactly as accurate as the market's, is bounded above by one, approached only by a trader who forecasts every outcome perfectly, and is unbounded below, because the denominator can be arbitrarily small.

3.3 Separating Luck from Skill

After computing BSS_j for every trader, we sift out luck to eliminate false prophets. A trader with no skill who buys only 97-cent favorites across twenty events goes without a loss with probability $0.97^{20} \approx 0.54$. Moreover, under our benchmark an uninformed taker is expected to score below zero, because crossing the spread places the taker's price further from a well-calibrated price under efficient markets (Appendix A.1). This property has a direct analogue in seasonal weather forecasting (Mason, 2004), with the difference that a meteorologist can always issue the reference forecast, whereas a taker cannot trade at the midpoint. Because we aim to identify the most clearly skilled traders, we test every trader active in at least 10 distinct events but keep only those active in at least 100, where the same 97-cent buyer would go undefeated with a probability of just 4.8 percent.

The test run randomizes the market outcomes each trader participated in. We obtain each trader's p-value following a framework similar to, but stricter than, that of Gómez-Cram, Guo, Jensen et al. (2026). In each of 1,000 draws, we redraw the outcome of every market a trader entered with probability equal to the weighted midpoint that trader faced there, averaged over their trades with each trade weighted by its contracts times the amount by which its price exceeds the midpoint, while holding prices, sizes, and market choices fixed. Markets in a nested event (a ladder of price thresholds) share one random draw per event so their outcomes stay consistent. All other markets are drawn independently.

Because we test hundreds of thousands of traders at once, a fixed threshold of $p \leq 0.05$ would flag a large number of them by chance. We therefore apply the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) at the 5 percent level across the restricted population. Ordering all traders' p-values as $p_{(1)} \leq \cdots \leq p_{(N)}$, we reject the null for every trader up to the largest k with $p_{(k)} \leq 0.05\,k/N$. The procedure bounds the expected share of false discoveries among flagged traders at 5 percent, the criterion that Barras, Scaillet, and Wermers (2010) advocate for performance evaluation in large cross-sections. Finally, we only keep traders with a positive score, meaning that they were more accurate than the market, and classify them as skilled.

3.4 Measuring Price Impact

Price Impact of Individual Orders. We ask how the market's forecast moves after a taker trades. An order is all trades by one taker in one market at one timestamp, and consecutive orders by the same taker on the same side with no other taker in between count as one. Let m_0 be the weighted midpoint just before the order, p the price paid on the YES scale, $D = +1$ if the taker bought YES and $D = -1$ if NO, and m_h the weighted midpoint just before a later order. Following Huang and Stoll (1996), the taker pays $D(p - m_0)$ over the midpoint, the midpoint then moves $D(m_h - m_0)$ in the direction of the trade, and the difference is what the maker keeps. We read m_h at the next order, at the first order between one and two minutes later, between five and ten minutes later, and at the last quote before close. Orders with no qualifying later quote drop out at that horizon. As a ceiling we report $D(Y - m_0)$, the outcome minus the midpoint on the side the taker bought. Averages weight each contract equally, and standard errors are clustered by event.

Order Flow and the Last Traded Price. Kalshi displays the last traded price, not the midpoint. Following Bartlett and O'Hara (2026), we therefore also estimate the regression of Glosten and Harris (1988) on five-minute intervals,

$$\Delta p_{jh} = \gamma_0 + \gamma_S Q^S_{jh} + \gamma_R Q^R_{jh} + \delta_S \Delta Q^S_{jh} + \delta_R \Delta Q^R_{jh} + \varepsilon_{jh}, \tag{4}$$

where Δp_{jh} is the change in market j's last traded price from the previous traded interval to interval h, and Q^S and Q^R are the net contracts bought on the YES side by skilled and other takers, in thousands. γ is the part of the price move that lasts and δ the part that reverses in the next interval. The coefficient γ is the λ of Kyle (1985), the lasting change in price per unit of net order flow, estimated here separately for skilled and other takers. We pool all markets and cluster standard errors by event.

3.5 Robustness Specifications

We report (3) with the quantity-weighted midpoint throughout. As a robustness check, we repeat the full procedure for the other three combinations of score and benchmark. The two scores are the ratio score (3) and an unnormalized, size-weighted mean of S_i (Appendix A.6). The two benchmarks are the quantity-weighted midpoint and the plain midpoint. Appendix B reports the results. In Appendix C we also compute scores out of sample, on markets not used to determine skill.

4 Who Are the Skilled Traders?

4.1 The Distribution of Skill and the Skilled Set

Figure 1 plots the distribution of BSS_j across scored accounts. The distribution is sharply concentrated in a narrow band around zero, with its mode just below zero and a longer left tail. Most takers add little information beyond the quoted price, and because crossing the spread costs accuracy relative to the midpoint, the typical uninformed taker scores slightly negative. The right tail, where skilled traders must sit, is thinner.

Table 1 shows how each step narrows the full trader population down to the skilled set. One third of accounts trade in at least 10 resolved events, and 571,621 trade in at least 100 distinct events, the population we deem to have enough observations. Of all accounts, 69,917 (0.72%, or 12.2%, of eligible accounts) clear the randomization test at a raw 5 percent threshold, against 28,581 expected by chance alone, and 14,694, roughly one in five, survive the false discovery rate correction with a positive score. The skilled set therefore comprises 0.15 percent of all accounts. The false discovery rate is controlled at 5 percent. The final skilled set is deliberately strict. It selects for consistent price-discovery ability, and it is neither a broad definition of skill nor a full list of skilled traders on the Exchange.

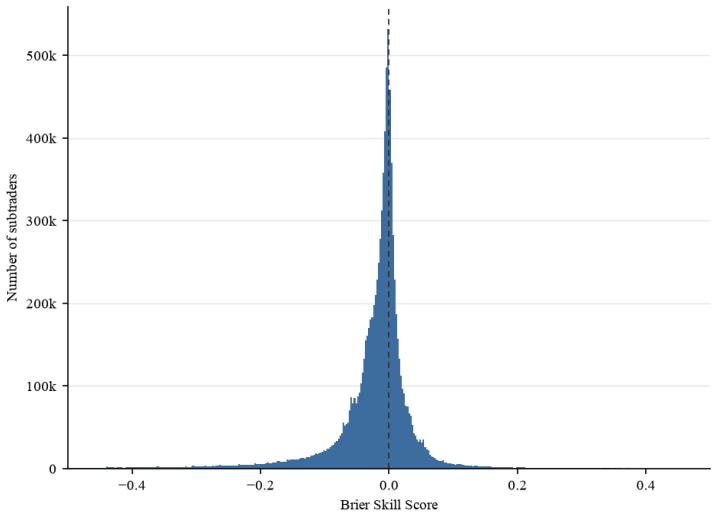


Figure 1: Distribution of the Brier skill score, BSS_j in equation (3), across all 9,656,615 scored accounts. The dashed vertical line marks zero, the score of a trader exactly as accurate as the quantity-weighted midpoint on the same trades. The horizontal axis is truncated at ±0.5.

Table 1: From all accounts to the skilled set

Stage	Criterion	N	% of all accounts
All accounts	Full sample	9,656,615	100.00
Active	≥ 10 distinct events	3,273,075	33.89
Eligible	≥ 100 distinct events	571,621	5.92
Raw significance	Randomization test, $p \leq 0.05$	69,917	0.72
Skilled	Benjamini–Hochberg at 5% and $\text{BSS}_j > 0$	14,694	0.15

Notes: The unit is the subaccount. *Eligible* accounts have participated in at least 100 resolved events and are tested in Section 3.3. *Raw significance* counts those with a raw $p \leq 0.05$ before any multiplicity adjustment. *Skilled* accounts additionally survive the Benjamini–Hochberg false discovery rate correction at the 5 percent level and have a positive score.

4.2 Concentration and Characteristics

Skilled traders account for 22.8 percent of transacted contract taker volume in the sample, roughly one hundred fifty times their 0.15 percent share of accounts. Within our skilled set, market coverage is fairly diversified. The trader with the largest footprint (percent of markets traded) trades in 22.5 percent of markets in the sample, more than double the next largest. This account is neither an institutional participant nor a designated market maker.

Figures 2 and 3 describe where skilled traders are registered and how long they have held their accounts. Skilled traders are concentrated in the most populous states, namely California, Texas, and New York. However, relative to the number of scored accounts, D.C., Wyoming, and Hawaii have the highest shares of skilled traders. The account-age distribution shows more pronounced spikes for all traders than for skilled ones. These spikes coincide with major events such as the U.S. presidential election of November 2024, Super Bowl LX (February 8, 2026), and

the 2026 FIFA World Cup (June 11 to July 19, 2026). This influx has also attracted more skilled accounts. The results show that skill does not require a long-tenured account, since a substantial share of skilled traders joined the platform within the past year.

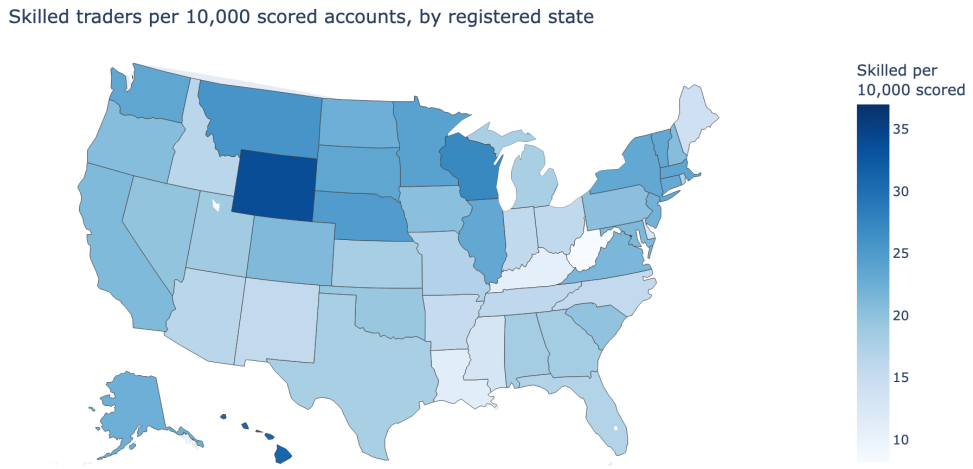


Figure 2: Skilled traders per 10,000 scored accounts, by registered state. Accounts opened through brokers, accounts with no profile, and non-US accounts are excluded.

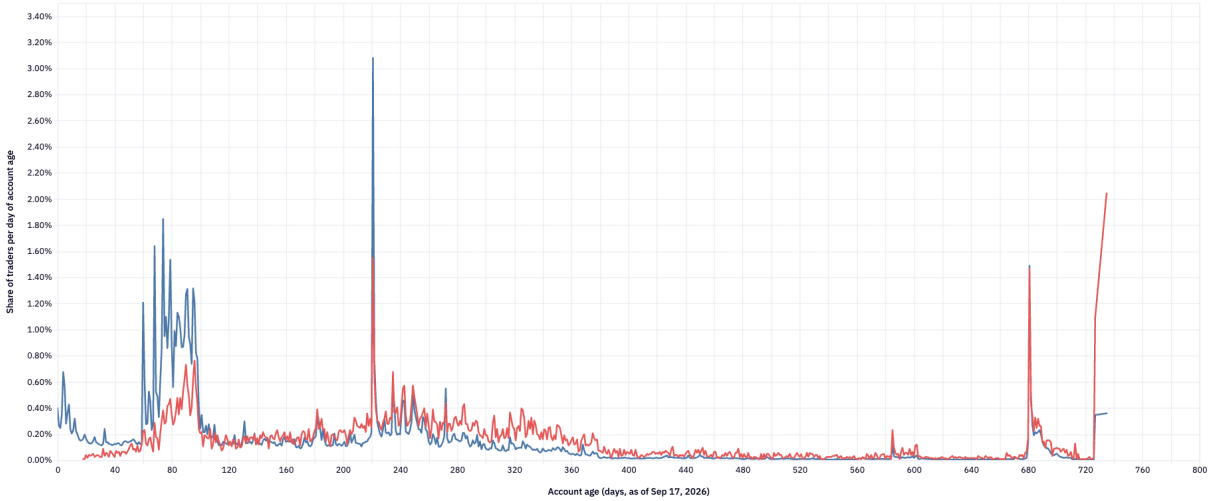
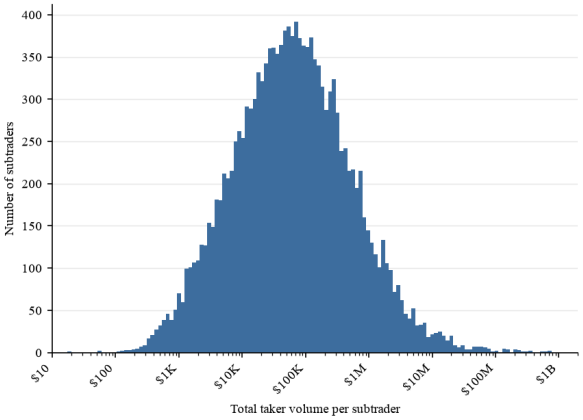


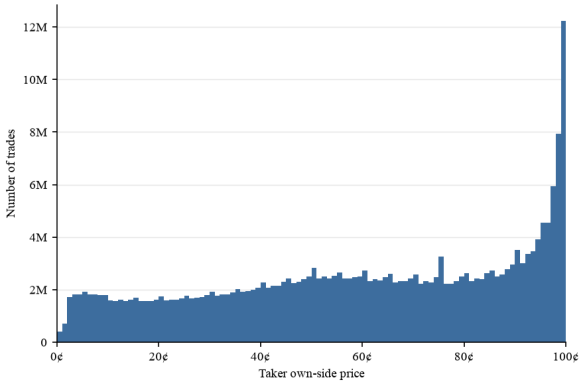
Figure 3: Account age, in days as of September 17, 2026. Each line reports the share of a group's accounts at each day of account age, for skilled traders (red) and for the full trader population (blue).

Figure 4 summarizes key patterns among skilled traders. Their total taker volume is approximately log-normally distributed and centered just below $100,000 (panel a). Gross taker profit and loss is centered just above zero and spreads over several orders of magnitude in both directions, with a longer tail of gains than of losses. Their prices sit closer to the eventual outcome than the weighted midpoint, but they tend to overshoot the true probability. Panel (b) shows that skilled taker trades are distributed across the full price range but concentrated at high own-side prices, with the single largest mass immediately below one dollar. In aggregate, skilled traders are found across the entire probability scale. However, panel (c) shows that the

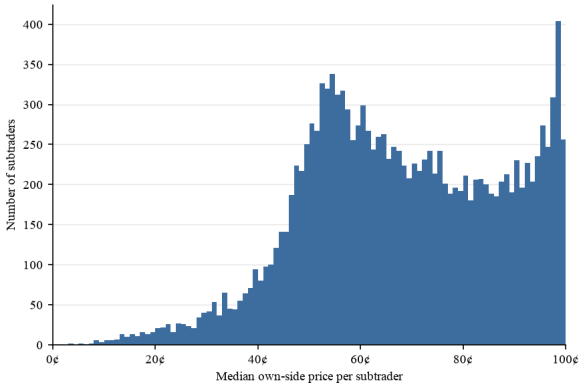
typical skilled trader skews toward trading at favorite prices. The median entry price per account forms a broad hump around 50 cents and a sharp spike near 99 cents. This may be because more markets trade near these two prices than elsewhere on the scale.



(a) Total taker volume per account



(b) Own-side price of every taker trade



(c) Median own-side entry price per account

Figure 4: Trading characteristics of the 14,694 skilled traders. Panels (a) and (c) count accounts; panel (b) counts trades. Volume is in dollars on a logarithmic scale; prices are stated on the trader's own side of the contract.

4.3 The Sources of Skill

To ask what skilled traders do better than the market, we apply the partition of Murphy (1973) to the Brier scores of each trader's prices and, separately, to those of the market's weighted midpoints prevailing on the same trades. For forecasts f_i, either the trader's prices p_i or the market's midpoints $m_{w,i}$, we sort trades into $K = 20$ bins of width 0.05 and weight by contracts throughout. Let W_k be the contracts in bin k and $W = \sum_k W_k$, and let $\bar{f}_k$, $\bar{o}_k$, and $\bar{o}$ be the contract-weighted mean forecast in bin k, mean outcome in bin k, and overall mean outcome.

The trade-level Brier score then satisfies:

$$\frac{1}{W}\sum_i w_i(f_i - Y_i)^2 = \underbrace{\sum_k \frac{W_k}{W}(\bar{f}_k - \bar{o}_k)^2}_{\text{REL}} - \underbrace{\sum_k \frac{W_k}{W}(\bar{o}_k - \bar{o})^2}_{\text{RES}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{UNC}} \tag{5}$$

$\sigma^2_{f,k}$ and $\sigma_{fY,k}$ are the contract-weighted variance of the forecast and its covariance with the outcome within bin k. *Reliability* (REL, lower is better) measures calibration, for example whether contracts bought near 90 cents paid off about 90 percent of the time. *Resolution* (RES, higher is better) rewards separating the events that happened from those that did not (i.e. rewards discernment), rather than quoting one hedged number everywhere. *Uncertainty* (UNC) measures how hard the events were to call and depends only on how often they occurred. Because we score trader and market on the same trades, uncertainty is identical for both and cancels between them.

An Illustrative Skilled Trader. Figure 5 and Table 2 apply the decomposition to a single example trader who clears the outcome randomization test but not the Benjamini–Hochberg adjustment. Suppose this trader takes the opposing side of a perceived favorite–longshot bias, and its trades lie entirely at own-side prices above 75 cents. Under that bias, favorites trade too cheaply (Snowberg and Wolfers, 2010; Bürgi, Deng, and Whelan, 2026). The contracts this trader bought paid off 88.9 percent of the time. Because a taker crosses the spread, the trader paid more than the weighted midpoint on every trade, and yet the trader's prices sat closer to the true outcomes. That can happen only if market prices were inaccurate at the moment of the trades. In the reliability diagram, calibration is the vertical distance of each point from the dashed 45-degree line, and resolution is the distance of those points from the horizontal base-rate line at 0.889. The trader beat the market on both. The trader's Brier score is 0.0832 against the market's 0.0844, and dividing the gap by the market's own Brier score gives a Brier skill score of 1.5 percent (lower squared error).

Table 2: Murphy decomposition, illustrative skilled trader against the market

	Trader	Market	Difference
Reliability (REL), lower is better	0.00577	0.00644	−0.00067
Resolution (RES), higher is better	0.02099	0.02007	0.00091
Uncertainty (UNC)	0.09870	0.09870	0.00000
Within-bin residual	−0.00033	−0.00065	0.00032
Brier score	0.08316	0.08442	−0.00126
Brier skill score			0.01497

Notes: The trader shown in Figure 5, decomposed as in equation (5) and compared with the weighted midpoints on the same trades; 20 bins of width 0.05, contract-weighted. The Brier score equals UNC − RES + REL plus the within-bin residual. The Brier skill score is the difference in Brier scores divided by the market's Brier score. Differences may not sum exactly because of rounding.

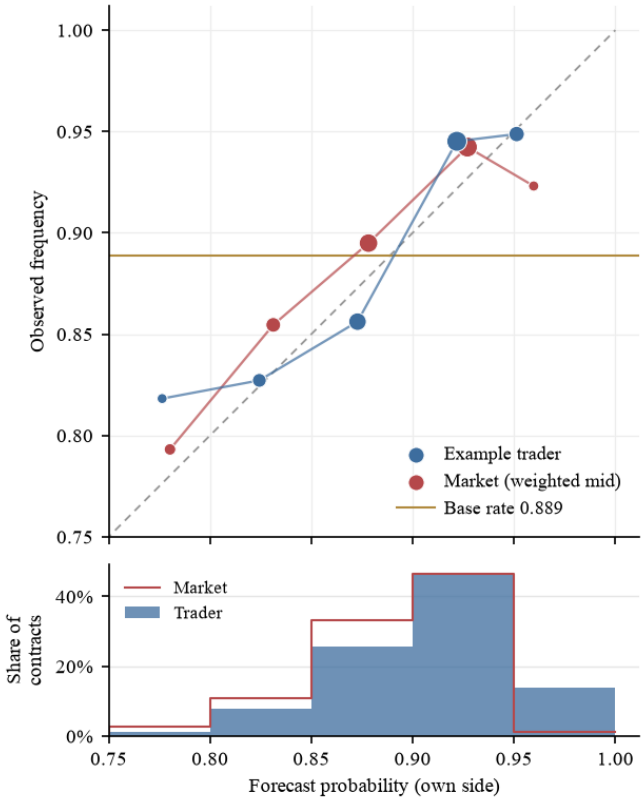


Figure 5: Reliability diagram for an illustrative skilled trader (blue) and for the weighted midpoints prevailing on the same trades (red). Marker size is proportional to contracts in the bin. The dashed line is perfect calibration and the horizontal line the contract-weighted base rate. The lower panel shows each forecast's share of contracts by bin. Because the trader crosses the spread, the trader's own-side prices sit above the midpoint on the same trades, which shifts the trader's distribution to the right.

The Skilled Population. Table 3 applies the same decomposition to each of the 14,694 skilled traders. Panel A reports the mean of each component across traders. On average, skilled traders have a lower Brier score on their trades of 0.13813, compared to the market's 0.14705, or about 6 percent of the market's own error. The larger part of this improvement comes from calibration. Reliability falls by 0.00570, roughly 64 percent of the total. Resolution rises by 0.00319, roughly 36 percent. The within-bin residual is negligible.

Panel B classifies each trader by whether their reliability is lower and their resolution higher than the market's on their own trades. Nearly half of skilled traders, 48.8 percent, are better on both dimensions. Another 23.3 percent are better on calibration only, and 27.4 percent on resolution only. Those better on both dimensions account for 86.5 percent of skilled-trader contract volume and have by far the highest median skill score, 0.0382, twice that of the skilled population as a whole. A small residual group of 66 traders, 0.45 percent, is better on neither dimension. Their positive scores arise from the within-bin term, and they account for 0.1 percent of contracts. The groups also differ in profit, which seems to be associated with scoring better on calibration.

Table 3: Murphy decomposition, skilled traders against the market

Panel A. Components, mean across traders

	Trader	Market	Difference
Reliability (REL), lower is better	0.00978	0.01548	−0.00570
Resolution (RES), higher is better	0.07237	0.06918	0.00319
Uncertainty (UNC)	0.20087	0.20087	0.00000
Within-bin residual	−0.00016	−0.00012	−0.00003
Brier score	0.13813	0.14705	−0.00892

Panel B. Where the edge comes from

	N	% of traders	% of contracts	Median BSS_j
Better calibration *and* resolution	7,176	48.84	86.49	0.0382
Better calibration only	3,425	23.31	7.33	0.0111
Better resolution only	4,027	27.41	6.06	0.0118
Better at neither	66	0.45	0.12	0.0017
All skilled traders	14,694	100.00	100.00	0.0189

Notes: The decomposition of equation (5) applied to each of the 14,694 skilled traders and to the weighted midpoints on the same trades; 20 bins of width 0.05, contract-weighted. Panel A reports the mean of each component across traders; the third column is the mean of the paired trader-minus-market difference. Each column sums to the Brier score. UNC is identical by construction, since trader and market are scored on the same trades and therefore share the same base rate. Panel B classifies each trader by whether their REL is below and their RES above the market's on their own trades. Percentages of contracts may not sum to 100 because of rounding.

Figure 6 compares how often the side each taker bought settled at \$1 with the price paid and with the weighted midpoint faced, for skilled takers and for all other takers. Skilled takers buy the side the midpoint underprices by about 4 percentage points, pay about 2 points over the midpoint, and keep about 2 points. Other takers trade at midpoints that are already calibrated, so the 1.2 points they pay over the midpoint is a loss. The two groups also differ in the side they buy. Other takers buy YES on 67 percent of their trades, a tilt that Becker (2026) links to a premium paid on YES contracts. Skilled takers buy YES on 53 percent. Because the skilled set is chosen on outcomes, part of the distance between the skilled curves and the diagonal is built in by selection. The persistence tests that follow address whether the skill is real.

4.4 Persistence

A skilled designation is informative about traders only if it persists through time. The mutual fund performance literature has long used persistence as a test of skill (Carhart, 1997). To check for persistence, we split events at random into two equal halves, a training half (P1) and a test half (P2), so that both halves span the full sample period. We then repeat the entire procedure separately in each half, including the 100-event restriction, the randomization test, and the false discovery correction. The panel consists of the 271,308 traders who trade in at least 100 distinct

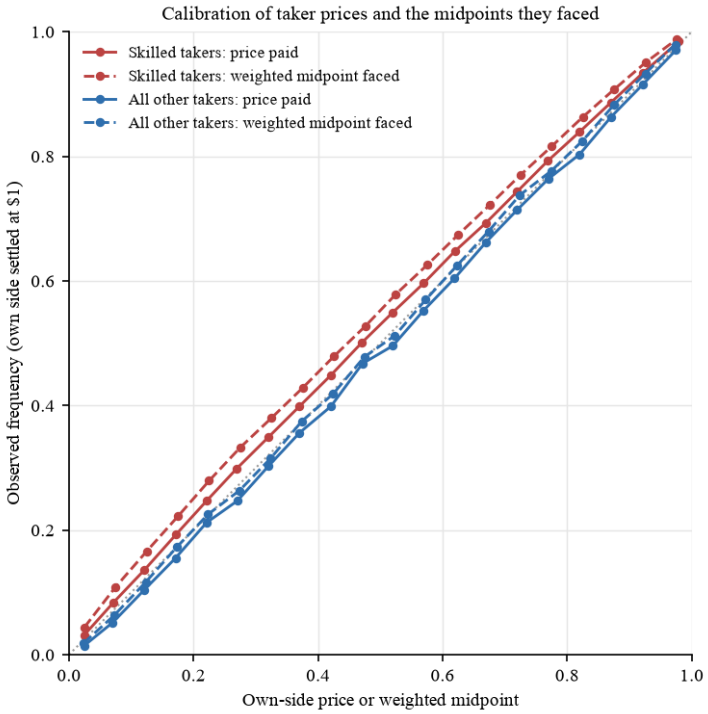


Figure 6: Observed frequency with which the taker's side settled at $1, by own-side price paid (solid) and by the weighted midpoint faced (dashed), for skilled takers (red) and all other takers (blue). Bins of width 0.05, contract-weighted. The dotted line is perfect calibration.

events in both halves.

Table 4 reports the results. Of the 8,197 panel traders skilled in the training half, 44.5 percent are skilled again in the test half, close to the rate that Gómez-Cram, Guo, Jensen et al. (2026) report. Among all panel traders, by contrast, only 3.0 percent are skilled in the test half. Because each half contains roughly half of every trader's events, the power to detect skill in either half is materially lower than in the full sample. As a result, fewer traders qualify as skilled in P1.

Table 4: Train–test persistence of the skilled designation

Group	N	Skilled in test half (P2)
Skilled in training half (P1)	8,197	44.5%
All panel traders	271,308	3.0%

Notes: Events are assigned independently at random to a training half (P1) or a test half (P2), so that both halves span the full sample period, and the full procedure is repeated independently in each half. The panel comprises traders who trade in at least 100 distinct events in both halves. The repeat rate is the share of each group that is also classified as skilled in P2.

5 Skilled Traders and Market Calibration

Section 4 establishes that skilled traders exist, trade heavily, and remain skilled out of sample. We now ask whether their presence is associated with more accurate markets. If it is, skilled traders would explain the aggregate calibration documented in prediction markets more precisely than volume or the total number of participants can. Throughout this section, the skilled designation is from the full sample. Appendix C reports a difference-in-differences design around skilled entry, including a version with skill computed out of sample.

5.1 Coverage

Table 5 reports the cumulative distribution of the number of skilled traders in each market, by category. Skilled traders are nearly ubiquitous. In the pooled sample, only 7.2 percent of markets receive no skilled trader at all, the median market has four, and 28.5 percent have more than ten. Coverage varies considerably across categories. It is densest in Mentions, where only 0.8 percent of markets lack a skilled trader, and in Elections and Politics. In all three, the median market has between six and ten. It is thinnest in Finance, where 55.4 percent of markets have at most one skilled trader. Mentions is also the category that Kagan and Baiocchi (2026) found hardest to forecast at longer horizons. The hardest-to-price markets thus draw the largest number of skilled traders. This pattern is anticipated in the literature, since the returns to information are greatest where prices are least informative (Grossman and Stiglitz (1980)).

Table 5: Cumulative percentage of markets with at most t skilled traders, by category

	Number of skilled traders in the market, t											
Category	0	≤1	≤2	≤3	≤4	≤5	≤10	≤20	≤30	≤40	≤50	≤100
Elections	2.8	12.5	22.5	31.8	41.8	49.5	74.8	89.7	94.3	96.4	97.5	99.3
Politics	4.4	13.2	21.2	28.9	36.8	44.2	72.1	90.1	94.1	95.9	96.9	99.1
Sports	8.2	26.0	37.5	45.7	51.9	56.7	71.2	83.6	89.4	92.7	94.7	98.2
Culture	6.2	38.1	52.2	61.1	68.0	73.1	87.0	95.6	98.2	99.0	99.4	99.8
Crypto	6.2	24.0	33.8	40.8	46.3	50.7	64.0	79.1	88.0	92.6	94.9	98.3
Commodities	5.9	33.7	48.1	58.8	67.0	73.4	88.6	94.7	97.7	99.1	99.7	100.0
Climate	4.5	27.1	37.1	43.6	48.8	53.2	70.0	86.7	94.5	98.0	99.4	100.0
Economics	9.2	32.6	48.2	59.2	67.8	74.3	91.0	98.4	99.6	99.9	99.9	100.0
Mentions	0.8	3.7	7.3	12.1	17.9	24.1	53.0	82.1	92.4	96.5	98.3	99.9
Finance	8.2	55.4	74.1	84.1	89.3	92.7	98.6	99.9	100.0	100.0	100.0	100.0
Tech & Science	6.6	23.7	37.0	48.5	57.2	64.2	85.1	95.6	98.5	99.4	99.7	100.0
Pooled	7.2	26.9	38.3	46.3	52.3	57.1	71.5	84.3	90.5	93.8	95.6	98.6

Notes: Each cell reports the percentage of resolved markets in the category in which at most t distinct skilled traders took liquidity. The first column is therefore the share of markets with no skilled trader.

5.2 When Skilled Traders Trade

Figure 7 compares when in a market's life skilled traders and all other traders place their volume. In Sports, both groups trade overwhelmingly in the final decile, when games are being played. That timing fits the reality that most information is revealed during the event. In several categories where information arrives more gradually, however, skilled traders are markedly earlier.

In Elections, Economics, and Culture, skilled traders place three to four times the share of their volume in a market's opening decile that other traders do, and they remain over-represented through the first half of market life. This pattern is consistent with skilled traders trading when prices are least informative, early in a market's life and before most other traders arrive.

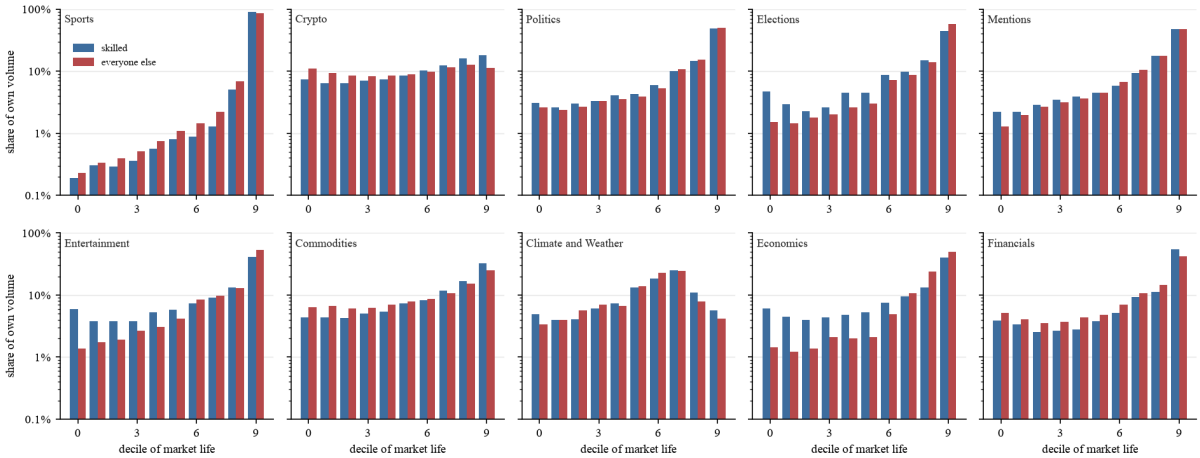


Figure 7: Share of each group's contract volume by decile of market life, within category. Blue bars are skilled traders and red bars all other traders. Decile 0 is the first tenth of a market's life, from listing to close, and decile 9 the last. Logarithmic vertical axis.

5.3 Markets with and without a Skilled Trader

We next compare markets with exactly one skilled trader to markets with none. Because skilled traders choose where to trade, an unconditional comparison would confound their effect with the characteristics of the markets they select. We therefore match each market with no skilled trader to every market in the same series that has exactly one, with total volume within 20 percent and lifespan within 10 percent. We then compare each zero-skilled market with the mean of its matches. Figure 8 reports the mean difference in Brier score between the zero-skilled and one-skilled markets at horizons from one day before close to close. A positive value means the market with a skilled trader was more accurate.

In the pooled sample, the two groups are close to indistinguishable a day before close and at intermediate horizons, with the market containing a skilled trader slightly less accurate four to twelve hours out. The gap opens in the final hours. By close, markets with a skilled trader have Brier scores roughly 0.027 lower than their matched counterparts. Some categories show the gap more consistently than the pooled result. In Culture, for example, the difference is close to or above 0.04 at every horizon. The positive difference is highest in Mentions, where it exceeds 0.05 from two hours before close. The pooled effect appears in the final hours, which matches Figure 7, where volume clusters in the final decile of market life. In several categories, so does information.

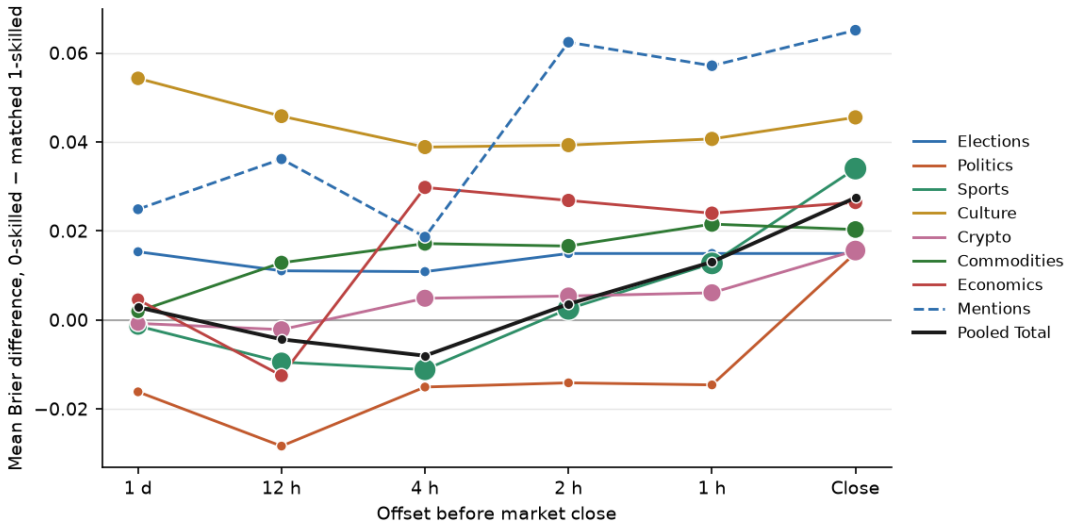


Figure 8: Mean difference in Brier score between markets with no skilled trader and matched markets with exactly one, by category and horizon before close. Markets are matched on exact series ticker, total volume within 20 percent, and lifespan within 10 percent. A positive value means the market with one skilled trader was more accurate. The black line pools all categories.

5.4 Prices and Accuracy after a Skilled Trader's Trade

We measure how far the weighted midpoint moves in the direction of each taker order, relative to the price paid over it (Huang and Stoll, 1996). Figure 9 follows the weighted midpoint after each taker order. Skilled takers buy the side that the pre-trade midpoint underprices by 3.8 cents on average, and they pay 1.5 to 1.8 cents over the midpoint. By the next order the midpoint has moved 1.7 cents their way, by one minute 2.8 cents, and by close 3.3 cents. The move does not reverse. Within a minute it exceeds what the skilled taker paid, so the maker on the other side loses 1.3 to 1.5 cents per contract. Other takers buy sides the midpoint already prices correctly, at −0.1 cents, and the midpoint moves 0.3 cents or less after their trades. Makers keep most of the 1.2 cents these takers pay.

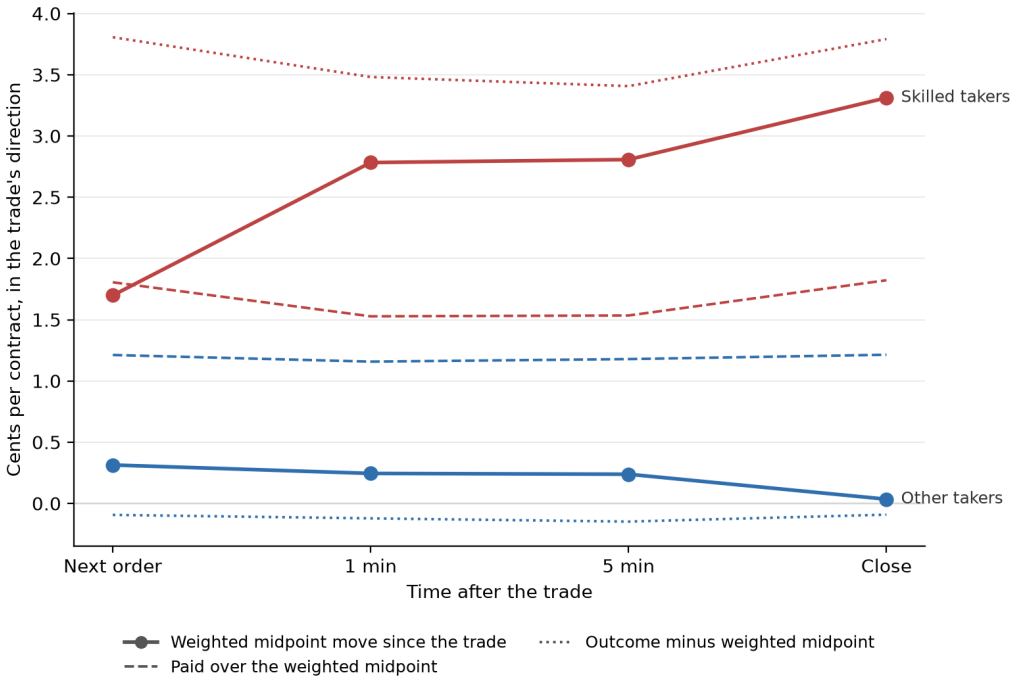


Figure 9: Weighted midpoint after a taker's trade, in cents per contract in the direction of the trade. Solid: how far the weighted midpoint has moved since the trade. Dashed: price paid over the pre-trade weighted midpoint. Dotted: outcome minus the pre-trade weighted midpoint, on the side bought. The gap between solid and dashed is what the maker keeps. Next order is the book when the next taker arrives; 1 and 5 min are the first order between one and two, or five and ten, minutes later; Close is the last quote before close. 100 million skilled and 526 million other taker orders in markets opened from October 2025.

The move is not only the mechanical effect of large orders on the book. When skilled takers place orders that take less than a quarter of the depth at the best price, the midpoint moves only 0.2 cents at first, yet within five minutes it has moved 1.1 cents their way. Orders by other takers that take the whole best level move it 0.8 cents, and that move fades to 0.2 cents by close. The later drift is consistent both with skilled traders moving the price and with their trading ahead of news.

We regress five-minute changes in the last traded price on net skilled and other order flow, separating the lasting from the reversing part (Glosten and Harris, 1988). Table 6 turns to the price Kalshi displays. A thousand contracts of net skilled buying move the last traded price by 0.107 cents, and the move lasts. The part that reverses is close to zero. Net buying by other takers has only a negligible lasting effect. A difference-in-differences design around skilled entry (Appendix C) points the same way. Against volume-matched markets, the Brier score of the weighted midpoint falls by 0.0012 to 0.0018 as each additional skilled trader enters, significant from the second to the tenth, with diminishing returns.

Table 6: Impact of net order flow on the last traded price

	Lasting (γ)	Temporary (δ)
Skilled takers	0.1069^{***} (0.0079)	0.0012 (0.0012)
Other takers	-0.0048^{***} (0.0005)	-0.0008 (0.0005)
Difference	0.1117^{***} (0.0082)	0.0020^{**} (0.0009)

Notes: Coefficients of equation (4), in cents per thousand contracts of net YES buying. Standard errors clustered by event in parentheses. Five-minute intervals, last traded price measured close to close; 90.6 million intervals in 530,000 events. The regression pools all markets without market fixed effects.

6 Limitations

Several features of the study limit the conclusions we can draw from it. First, the price-impact results are associations. After a skilled taker's trade the weighted midpoint moves toward the outcome and stays there, but part of the later move may reflect skilled traders trading ahead of news rather than moving the price. The skilled set is also chosen on outcomes, so the move to close is partly built in by selection. The difference-in-differences designs in Appendix C match controls on series, lifespan, and pre-entry volume, but not on unobserved market features that may both attract skilled traders and make pricing easier. They also assume that treated and control markets would have followed parallel paths absent entry, an assumption that cannot be tested directly.

Second, the score does not separate trading behaviors that give it different meanings. Although we designed BSS_j to measure forecasting skill more narrowly than profit does, it also rewards pure arbitrageurs. Takers using algorithms that close arbitrage opportunities between underlyings (Commodities, Crypto, ...) and the event contract can score as skilled without forecasting in the sense of Tetlock and Gardner (2015). A future study that aims to measure forecasting skill alone could exclude Sports, Crypto and other short-term markets that trade on an underlying. We likewise keep taker trades that close an earlier position. A closing trade is still made at a chosen price against the prevailing weighted midpoint and so carries the same price-discovery content as an opening one. More broadly, the field would benefit from a definition of forecasting that separates its different kinds. Whether exploiting a documented, public pattern such as the favorite–longshot bias counts as forecasting skill remains a valid and open question.

Third, the false discovery rate correction controls the expected share of false discoveries at 5 percent, so some of the skilled accounts may be misclassified. The bound also assumes that traders' p-values are independent or positively dependent. Traders in the same events share the same randomized outcomes, and those who take opposite sides of a market have negatively dependent p-values. Aggregate statistics for the skilled set are robust to misclassification of this size, but statements about individual accounts are not.

Fourth, the sample covers a single year, during which spreads narrowed markedly and the trader base grew. Future work could test whether the size and composition of the skilled set are stable across market conditions, and in particular whether the set grows as trading costs fall.

7 Conclusion

Volume and the number of participants are often cited as simple measures of a market's information content. Past literature, including our own work, showed that volume and the number of traders are both associated with better market accuracy and calibration, and that calibration also improves closer to resolution (Kagan and Baiocchi, 2026). However, these are aggregate market properties. The market microstructure literature has long distinguished between types of participants. Skilled presence complements those aggregate properties and is informative beyond volume. With series, volume, and lifespan held fixed, markets with a skilled trader close more accurately (Figure 8), and after a skilled taker's trade the weighted midpoint moves toward the outcome at once, a move that is not reversed (Figure 9). Against volume-matched controls, the midpoint moves toward the outcome as each additional skilled trader enters a market, with diminishing returns (Appendix C). Skilled-trader coverage could therefore join volume and time to resolution as a market-level indicator of how far a price can be trusted. Measuring the share of skilled relative to noise traders would likewise complement these findings.

The crowd still matters, partly because it gives skilled traders a reason to trade. Noisy order flow raises the return to information (Grossman and Stiglitz, 1980), and compensates the makers facing skilled takers (Bartlett and O'Hara, 2026), so the trader count, uninformed participants included, complements skilled presence. Skill also works together with time to resolution. Skilled traders enter early where information arrives gradually, and late in Sports (Figure 7). Separating skilled traders from the crowd therefore points to an explanation for a puzzle in the companion study, Kagan and Baiocchi (2026). There, markets with the most traders were, at the longest horizon, very slightly worse calibrated than markets with the fewest, which suggests that the raw count of participants is an imperfect proxy for what matters.

Skilled takers pay the spread to move prices toward the truth. For the median skilled trader, the edge is small, about 2 percent of the market's error, and comes mostly from calibration (Section 4.3). Scoring better on calibration is also associated with higher profitability than scoring better on resolution alone. Overall, prices become more accurate after skilled traders' trades, whether or not the traders profit.

Two implications follow for the durability of calibration as prediction markets grow. The first is reassuring. Calibration holds even when the average participant is poorly informed, provided a well-calibrated minority is present and the market lets that minority trade cheaply. The narrowing of spreads lowers the hurdle a skilled trader's edge must clear, and so should, if anything, draw more skilled trading into the market. The second is a caution about market dynamics. The mechanism documented here is weakest in markets that attract no skilled trader, 7.2 percent of our sample, and in categories where skilled coverage is thin.

To our knowledge, this is the first study to identify skilled traders at this scale from the account-level record of a U.S.-regulated exchange and to link them directly to market-level calibration. The natural next step is to understand, on one hand, the personal characteristics of those identified traders, and, on the other, how the size of the skilled set responds to trading costs, market design, and the growth of the platform.

References

Akey, Pat, Vincent Grégoire, Nicolas Harvie, and Charles Martineau. 2026. "Who Wins and Who Loses in Prediction Markets? Evidence from Polymarket." CEPR Discussion Paper 21615; SSRN Working Paper 6443103.

Barras, Laurent, Olivier Scaillet, and Russ Wermers. 2010. "False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas." *Journal of Finance* 65 (1): 179–216.

Bartlett, Robert, and Maureen O'Hara. 2026. "Adverse Selection in Prediction Markets: Evidence from Kalshi." Rock Center for Corporate Governance at Stanford University Working Paper 266; SSRN Working Paper 6615739.

Becker, Jonathan. 2026. "The Microstructure of Wealth Transfer in Prediction Markets." SSRN Working Paper 7217640.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57 (1): 289–300.

Berg, Joyce E., Forrest D. Nelson, and Thomas A. Rietz. 2008. "Prediction Market Accuracy in the Long Run." *International Journal of Forecasting* 24 (2): 285–300.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Bürgi, Constantin, Wanying Deng, and Karl Whelan. 2026. "Makers and Takers: The Economics of the Kalshi Prediction Market." CESifo Working Paper 12122.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2): 238–249.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs." *Quarterly Journal of Economics* 134 (3): 1405–1454.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–2996.

Della Vedova, Joshua. 2026. "Who Profits from Prediction? Execution, not Information." University of San Diego Working Paper; SSRN Working Paper 6191618.

Desai, Maanit, Yirun Zhang, Ryan Holbrook, Kait O'Neil, and Maggie Demkin. 2024. *Jane Street Real-Time Market Data Forecasting*. Kaggle competition.

Diercks, Anthony M., Jared Dean Katz, and Jonathan H. Wright. 2026. "Kalshi and the Rise of Macro Markets." Finance and Economics Discussion Series 2026-010. Washington, DC: Board of Governors of the Federal Reserve System.

Fama, Eugene F., and Kenneth R. French. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65 (5): 1915–1947.

Glahn, Harry R., and Dale L. Jorgensen. 1970. "Climatological Aspects of the Brier P-Score." *Monthly Weather Review* 98 (2): 136–141.

Glosten, Lawrence R., and Lawrence E. Harris. 1988. "Estimating the Components of the Bid/Ask Spread." *Journal of Financial Economics* 21 (1): 123–142.

Glosten, Lawrence R., and Paul R. Milgrom. 1985. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14 (1): 71–100.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Gómez-Cram, Roberto, Yunhan Guo, Theis Ingerslev Jensen, and Howard Kung. 2026. "Prediction Market Accuracy: Crowd Wisdom or Informed Minority?" SSRN Working Paper 6617059.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–277.

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70 (3): 393–408.

Gu, A., N. Kagan, A. Sun, J. Wu, and H. Xu. 2026. "When Do Prophets Profit in Prediction Markets?" arXiv:2607.06166.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64 (4): 605–654.

Huang, Roger D., and Hans R. Stoll. 1996. "Dealer versus Auction Markets: A Paired Comparison of Execution Costs on NASDAQ and the NYSE." *Journal of Financial Economics* 41 (3): 313–357.

Jensen, Michael C. 1968. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance* 23 (2): 389–416.

Kagan, Nicole, and Rubens Baiocchi. 2026. "Calibration in Prediction Markets: Theory and Evidence." SSRN Working Paper, August 20.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica* 53 (6): 1315–1335.

Luong, K. L., and G. Heesen. 2026. "The Wisdom of the Few: Skilled Traders and Prediction Market Accuracy." SSRN Working Paper 6758662.

Mason, Simon J. 2004. "On Using 'Climatology' as a Reference Strategy in the Brier and Ranked Probability Skill Scores." *Monthly Weather Review* 132 (7): 1891–1895.

Murphy, Allan H. 1973. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology* 12 (4): 595–600.

Ng, Hunter, Lin Peng, Yubo Tao, and Dexin Zhou. 2026. "Price Discovery and Trading in Modern Prediction Markets." SSRN Working Paper 5331995.

Oliven, Kenneth, and Thomas A. Rietz. 2004. "Suckers Are Born but Markets Are Made: Individual Rationality, Arbitrage, and Market Efficiency on an Electronic Futures Market." *Management Science* 50 (3): 336–351.

Page, Lionel, and Robert T. Clemen. 2013. "Do Prediction Markets Produce Well-Calibrated Probability Forecasts?" *Economic Journal* 123 (568): 491–513.

Reichenbach, Felix, and Martin Walther. 2026. "Exploring Decentralized Prediction Markets: Accuracy, Skill, and Bias on Polymarket." SSRN Working Paper 5910522.

Snowberg, Erik, and Justin Wolfers. 2010. "Explaining the Favorite–Longshot Bias: Is It Risk-Love or Misperceptions?" *Journal of Political Economy* 118 (4): 723–746.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz. 2013. "Prediction Markets for Economic Forecasting." In *Handbook of Economic Forecasting*, vol. 2A, edited by Graham Elliott and Allan Timmermann, 657–687. Amsterdam: Elsevier.

Stoikov, Sasha. 2018. "The Micro-Price: A High-Frequency Estimator of Future Prices." *Quantitative*

Finance 18 (12): 1959–1966.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–199.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York: Crown.

Wilks, Daniel S. 2011. *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Oxford: Academic Press.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives* 18 (2): 107–126.

Wolfers, Justin, and Eric Zitzewitz. 2006. "Interpreting Prediction Market Prices as Probabilities." NBER Working Paper 12200.

A Properties of the Skill Measure

This appendix collects the formal properties of the per-trade accuracy gain S_i in equation (2) that motivate the design choices in Section 3. S_i is the numerator of BSS_j in equation (3). Throughout, prices are on the trader's own side, m is the benchmark midpoint, p the execution price, $\Delta = p - m$ the taker's displacement from the midpoint (the half-spread for a trade at the touch), $q = \Pr(Y = 1)$ the unobserved true probability, and $\delta = q - m$ the market's error.

A.1 Expected Score and the No-Skill Null

For any forecast x, $\mathbb{E}[(x - Y)^2] = (x - q)^2 + q(1 - q)$. The expected per-trade gain is therefore

$$\mathbb{E}[S] \;=\; (m - q)^2 - (p - q)^2 \;=\; 2\delta\Delta - \Delta^2. \tag{6}$$

Under the calibration null $\mathbb{E}[Y \mid m, p] = m$, so that $\delta = 0$, equation (6) gives $\mathbb{E}[S] = -\Delta^2 < 0$ for any taker who crosses the midpoint. Zero is therefore not the no-skill value: an uninformed taker pays the half-spread in accuracy as well as in money. This is why the randomization null of Section 3.3 sits below zero and why we additionally require a positive score for a trader to be classified as skilled.

A skilled taker crosses the spread only when the price is below their valuation, $q > p$, or equivalently

$$q > p \iff \delta > \Delta. \tag{7}$$

Every trade a truly skilled taker makes therefore lies in the region $\Delta \in (0, \delta)$, over which $\mathbb{E}[S]$ is positive and increasing in Δ. For a given edge, a trade that crosses a wider spread scores higher. Under a price-discovery reading this is the correct ordering: moving a thin market four cents toward the truth is more discovery than moving a liquid market half a cent.

A.2 A Worked Example: Accuracy Is Not Profit

Because $\mathbb{E}[S]$ rises and then falls in Δ, while expected profit per contract, $q - p = \delta - \Delta$, falls monotonically, the score is not monotone in profit. Fix the midpoint at $m = 0.30$ and the true probability at $q = 0.36$, so that $\delta = 0.06$ and $\mathbb{E}[S] = 0.0036 - (p - 0.36)^2$:

Trader	p	Δ	$q - p$	$\mathbb{E}[S]$
A	0.31	0.01	+0.05	0.0011
B	0.36	0.06	0.00	0.0036
C	0.40	0.10	−0.04	0.0020

Trader A buys at 0.31 and stops. A earns the most per contract but scores the least, because A never moves the price to q; A may have been too conservative, or may have faced a liquidity constraint. Trader B buys up the order book until the price reaches the true probability of 0.36, earning nothing on the marginal contract but achieving the maximum score. Trader C loses money and yet outscores A, because C's price is closer to q. C cannot, however, be a skilled trader: C's trade requires $\Delta > \delta$, a trade the trader believes to have negative value, which by (7)

rational informed play never enters. The non-monotone branch of the score therefore lies outside the region in which skilled takers operate.

A.3 Why Only Takers Are Scored

When a maker's resting order executes, the maker trades on the passive side of the midpoint, so that $\Delta < 0$. By equation (6), a maker who is correct about the direction of the market, $\delta > 0$, then has $\mathbb{E}[S] = 2\delta\Delta - \Delta^2 < 0$, and scores worse the better the price. The per-trade gain measures the information a trader adds by moving away from the quoted price; a maker who trades passively has crossed no distance from the quote to be judged on, and scoring makers on the same basis would penalize precisely the behavior that earns them the spread.

A.4 Choice of Benchmark: Weighted versus Plain Midpoint

The plain midpoint $m = (P^{\text{bid}} + P^{\text{ask}})/2$ ignores the quantities resting at the touch. With spread $s = P^{\text{ask}} - P^{\text{bid}}$ and order-book imbalance $I = Q^{\text{bid}}/(Q^{\text{bid}} + Q^{\text{ask}})$, the quantity-weighted midpoint of equation (1) satisfies $m_w = P^{\text{bid}} + sI$, so that

$$m_w - m \;=\; s\left(I - \tfrac{1}{2}\right). \tag{8}$$

All results of Appendix A.1 hold with m_w in place of m. What changes is the null. Suppose the book-implied expectation is m_w, so that an uninformed taker's trades have $q = m_w$. Scoring a taker who lifts the ask, $p = m + s/2$, against the plain midpoint then gives

$$\mathbb{E}\left[(m - Y)^2 - (p - Y)^2\right] \;=\; s^2\left(I - \tfrac{3}{4}\right), \tag{9}$$

which is positive whenever $I > 3/4$. Against the plain midpoint, an uninformed trader who buys into a book stacked three-to-one on the bid would record positive skill. Order-book imbalance is one of the mechanisms this paper seeks to separate from information, which is why the quantity-weighted midpoint is the primary benchmark. Because the two benchmarks differ only through this term, the gap $(m - Y)^2 - (m_w - Y)^2$ does not depend on the taker's execution price, and the difference between the numerators of a trader's scores under the two benchmarks measures how much of their apparent skill came from trading with imbalance.

A.5 Round Trips

Because each trade is scored at its own price, positions that are opened and closed before resolution are scored on their price path rather than on the outcome. Consider a trader who buys at $m_{\text{in}} + \Delta$ and later sells the same contract, equivalently buying the opposite side, at an own-side price of $(1 - m_{\text{out}}) + \Delta$, with equal half-spreads on both legs. The outcome cancels between the two legs, and their per-trade gains sum to

$$S_{\text{in}} + S_{\text{out}} \;=\; 2\left[\Delta\,(m_{\text{out}} - m_{\text{in}}) - \Delta^2\right], \tag{10}$$

a function of the midpoint path alone. Closed positions are therefore scored on short-horizon price forecasting, rewarding a trader who buys before the midpoint rises, and held positions on

forecasting the resolution. Both are valid observations of skill.

A.6 Normalization and Weights

The primary score (3) divides the size-weighted sum of per-trade gains by the market's own error on the trader's trades. The robustness specifications replace it with the unnormalized, size-weighted mean

$$\overline{S}_j \;=\; \frac{\sum_{i\in\mathcal{T}_j} w_i\, S_i}{\sum_{i\in\mathcal{T}_j} w_i}. \tag{11}$$

The two forms share a numerator and therefore always agree in sign. They differ in scale. Under calibration, the denominator of (3) has expectation $m(1-m)$ per contract, which falls from 0.25 at even-money prices to about 0.03 near 97 cents, so that the ratio form scales up a given absolute improvement for traders who concentrate on favorites or longshots. Reporting both forms ensures that the skilled set is not an artifact of that scaling.

Weights are contracts rather than capital at risk. A trade is a single object with two counterparties, and weighting by $w_i p_i$ would count the favorite side of the same transaction many times more heavily than the longshot side, reintroducing the dependence on the mix of prices that the score exists to remove.

B Randomization Results across Scores and Benchmarks

This appendix reports the robustness specifications of Section 3.

B.1 Scores and Benchmarks

The full procedure of Section 3.3, from scoring through the randomization test and the false discovery rate correction, is repeated for each of the two scores, the ratio BSS_j of equation (3) and the unnormalized mean $\overline{S}_j$ of equation (11), each measured against the two benchmarks, the quantity-weighted midpoint m_w of equation (1) and the plain midpoint m of Appendix A.4. We write $\mathrm{BSS}_j(m_w)$ and $\overline{S}_j(m_w)$ for the scores as defined in the main text, and $\mathrm{BSS}_j(m)$ and $\overline{S}_j(m)$ for the same scores with m_i in place of $m_{w,i}$ throughout. Under the plain benchmark, each market's outcome is redrawn with probability equal to the plain midpoint the trader faced there. $\mathrm{BSS}_j(m_w)$ is the score used throughout the paper, and its row of Table 7 reproduces Table 1.

Table 7 traces the eligible population to the skilled set under each combination. Three of the four produce skilled sets of similar size: 14,694 traders under $\mathrm{BSS}_j(m_w)$, 11,660 under $\overline{S}_j(m_w)$, and 11,358 under $\mathrm{BSS}_j(m)$. Under every combination, the number of traders that clear the raw 5 percent threshold is well above the 28,581 expected by chance alone.

B.2 The Unnormalized Score against the Plain Midpoint

Under $\overline{S}_j(m)$, no trader survives the correction, although 49,740 clear the raw threshold, about 1.7 times the number expected by chance. At this scale, the correction behaves as a threshold. With 1,000 draws the smallest attainable p-value is 1/1001, so on a population of 571,621 the Benjamini–Hochberg procedure rejects any trader only if about 2 percent of the population attain

Table 7: From eligible traders to the skilled set, by score and benchmark

Score	Midpoint	Eligible	Raw $p \leq 0.05$	Expected by chance	Skilled	% of all accounts	Median score, skilled
BSS_j	Weighted, m_w	571,621	69,917	28,581	14,694	0.15	0.0189
BSS_j	Plain, m	571,621	56,512	28,581	11,358	0.11	0.0119
$\overline{S}_j$	Weighted, m_w	571,621	60,998	28,581	11,660	0.12	0.0031
$\overline{S}_j$	Plain, m	571,621	49,740	28,581	0	0.00	–

Notes: The unit is the subaccount. Each row repeats the randomization test of Section 3.3 and the Benjamini–Hochberg correction on the same eligible population, for the score and benchmark midpoint shown. m_w is the quantity-weighted midpoint of equation (1) and m the plain midpoint. *Eligible* accounts trade in at least 100 distinct events. *Raw* $p \leq 0.05$ counts those that clear the randomization test before any multiplicity adjustment, and *Expected by chance* is the number that would clear it if no eligible account had any skill, 5 percent of the eligible population. *Skilled* accounts additionally survive the Benjamini–Hochberg correction at the 5 percent level and have a positive score. Median scores are in each score's own units. The first row is the skilled set used throughout the paper (Table 1).

it, and otherwise rejects none. The mean score against the plain midpoint falls short of that threshold because, for these takers, the plain midpoint is the harder benchmark. Takers mostly trade into the deeper side of the book, where the plain midpoint places a higher probability on the side taken than the weighted midpoint does, and on the eligible traders' trades its Brier score is lower, 0.1614 against 0.1619. The ratio form, which scales each trader's gain by the market's own error on the same trades, retains enough traders under the same benchmark to clear the threshold.

B.3 Overlap of the Skilled Sets

Table 8 reports how far the skilled sets agree. The primary set is the largest and contains most of the others: 96.7 percent of the traders skilled under $\overline{S}_j(m_w)$ and 78.0 percent of those skilled under $\mathrm{BSS}_j(m)$ are also skilled under $\mathrm{BSS}_j(m_w)$. Changing the score while keeping the benchmark moves the set less than changing the benchmark while keeping the score. In total, 7,648 traders are skilled under all three non-empty combinations, out of 17,532 skilled under at least one.

Table 8: Overlap of the skilled sets across scores and benchmarks

		% also skilled under			
Skilled under	N	$\mathrm{BSS}_j(m_w)$	$\mathrm{BSS}_j(m)$	$\overline{S}_j(m_w)$	$\overline{S}_j(m)$
$\mathrm{BSS}_j(m_w)$	14,694	100.0	60.3	76.7	0.0
$\mathrm{BSS}_j(m)$	11,358	78.0	100.0	67.7	0.0
$\overline{S}_j(m_w)$	11,660	96.7	66.0	100.0	0.0
$\overline{S}_j(m)$	0	–	–	–	–

Notes: Each row reports, for the traders skilled under the row's score and benchmark, the percentage who are also skilled under the column's. N is the size of the row's skilled set (Table 7). The matrix is not symmetric because the sets differ in size. The last row is undefined because the skilled set under $\overline{S}_j(m)$ is empty.

C Difference-in-Differences around Skilled Entry

We ask whether accuracy improves when a skilled trader enters a market. We use a difference-in-differences design to compare the change in each market's accuracy from before to after the entry of a skilled trader with the change around a matched point in similar markets that saw no new skilled entry. Because markets can have several skilled traders, we run this comparison separately for the first skilled entrant, the second, and so on up to the tenth, and call each of these a rung k. Each treated market and its matched controls form a two-by-two comparison, and we stack these comparisons as in Cengiz, Dube, Lindner et al. (2019). Following the logic of Callaway and Sant'Anna (2021), no control has already received the same entrant. This avoids the bias of a pooled two-way fixed-effects regression, which would use already-treated markets as controls (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021).

Treatment is defined by the skilled set of Section 4, estimated on the full sample. At rung k, a treated market is one whose k-th skilled taker places a first trade at time T. Its controls are markets with exactly $k-1$ skilled takers. We measure the instantaneous and cumulative effects on accuracy separately. For the instantaneous effect at entry, controls also include markets not yet treated whose own k-th entry comes after the treatment. Each control comes from the same series as the treated market but a different event, and must have a lifespan within 10 percent of the treated market's. Because a control has no treatment, we assign it a placebo τ at the same fraction of its window at which T falls in the treated market's, and take its first trade after τ, which must fall within 1 percent of the window. Matching on pre-treatment observables, we keep all controls whose contract volume from open to τ lies within 20 percent of the treated market's volume from open to T. Both need at least one trade before the treatment. The estimating equation, for market m in period t before or after the split, is

$$B_{mt} \;=\; \alpha_m \;+\; \beta\,\mathrm{Post}_{mt} \;+\; \gamma_k\,(\mathrm{Post}_{mt} \times \mathrm{Treated}_m) \;+\; \varepsilon_{mt} \tag{12}$$

$B_{mt} = (p_{mt} - Y)^2$ is the Brier score of price p_{mt} against the market outcome and α_m is a market fixed effect. For the instantaneous effect, p_{mt} is the price of the last trade before the treatment, and p_{mt} after it is the price of the treatment trade itself. An alternative version uses the quantity-weighted midpoints on either side of the treatment trade. For the cumulative effect, treated markets must close with exactly k skilled takers and controls with exactly $k-1$, and B_{mt} is the time-weighted mean Brier score from the $(k-1)$-th entry to the treatment and from the treatment to close. We estimate (12) by weighted least squares, giving each treated market a weight of one and each of its n_c control pairs a weight of $1/n_c$, so that γ_k is the average treatment effect on the treated at rung k. A negative γ_k means that, relative to its controls, the market became more accurate after the k-th skilled trader entered. We cluster standard errors two ways (Cameron, Gelbach, and Miller, 2011), by series and by the account of the k-th skilled trader.

Figure 10 reports the change in Brier score at the entry of the k-th skilled trader. On the quantity-weighted midpoint, the Brier score falls by 0.0012 to 0.0018 relative to controls, significant from the second rung onward, with diminishing returns. On the last traded price the estimate is larger, 0.0061 at the first rung and 0.0079 at the second, decaying to 0.0028 at the tenth. Because the post-entry price on this measure is the skilled trader's own fill, it is an upper bound on how far the market reprices.

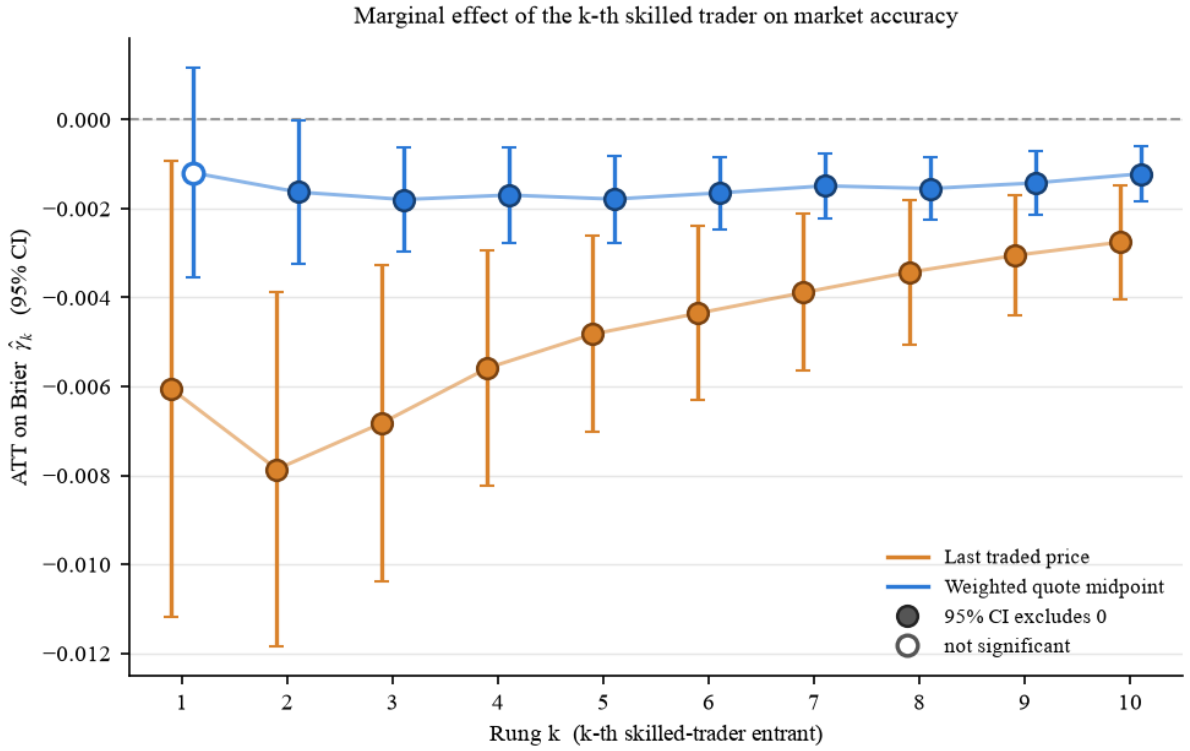
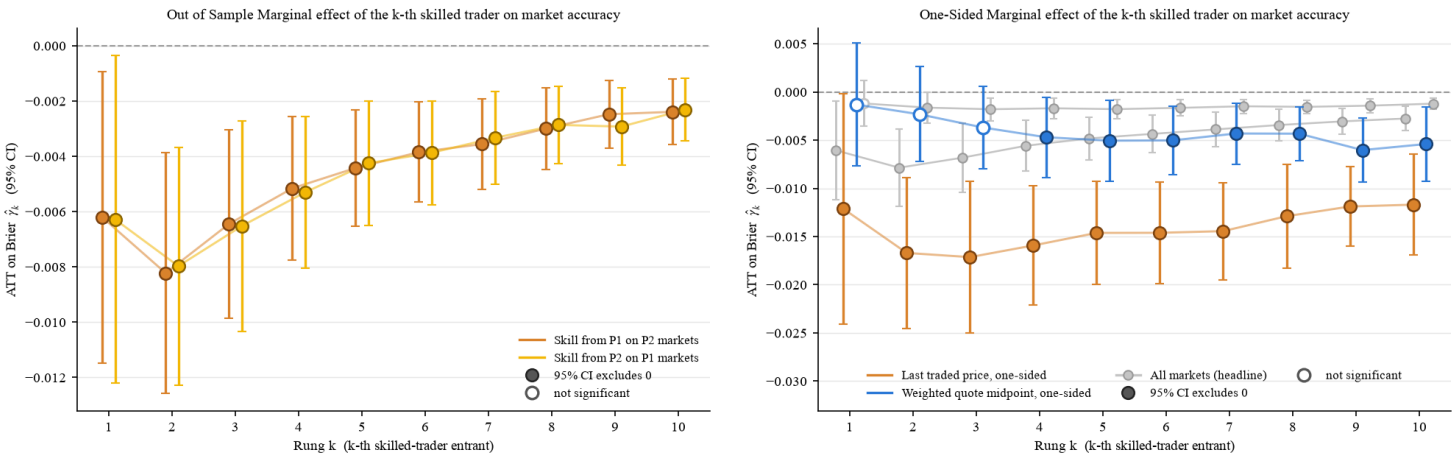


Figure 10: Change in the Brier score at the entry of the k-th skilled trader, relative to controls. Bars are 95 percent confidence intervals. Filled markers exclude zero.

Figure 11(a) classifies traders on one half of the sample and measures their effect on markets in the other. The estimates barely move, at 0.006 for the first skilled trader and 0.008 for the second, and all ten rungs remain significant in both directions. The effect is therefore not an artifact of identifying skilled traders in the same markets where we measure their impact. Panel (b) keeps only one-sided markets, in which no skilled trader took both sides. There the effect on the traded price is two to four times the headline estimate, between 0.012 and 0.017, and the midpoint effect is significant from the fourth rung onward, between 0.004 and 0.006.



(a) Out-of-sample skill

(b) One-sided markets

Figure 11: Change in the Brier score at entry. Gray markers in (b) repeat Figure 10.

Figure 12 compares the time-weighted Brier score from entry to close of markets that end with exactly k skilled traders against those that end with $k-1$. Because markets are selected on their skilled count at close, this comparison is descriptive. The first rung is too noisy to read, at +0.004 with a standard error of 0.011. From the second rung onward every estimate is a reduction, of 0.004 to 0.009, significant at the second and at the sixth through tenth. On the midpoint, the cumulative estimate is not distinguishable from zero at any rung.

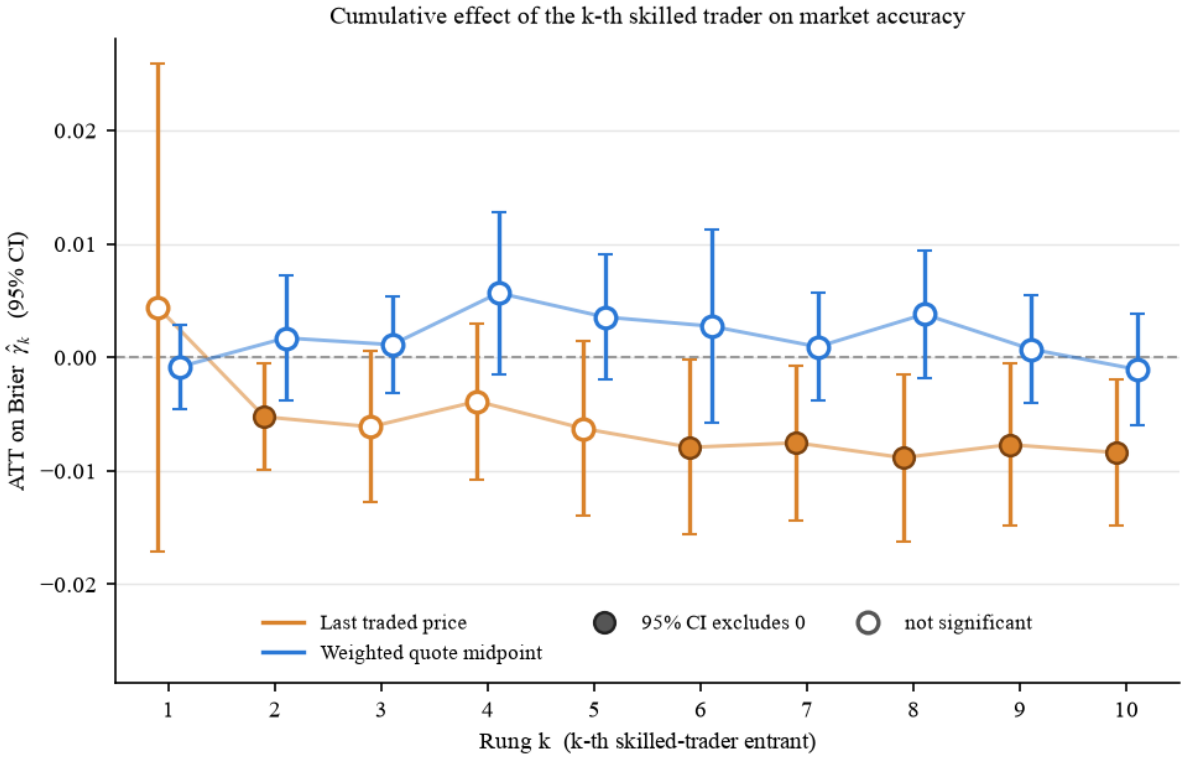
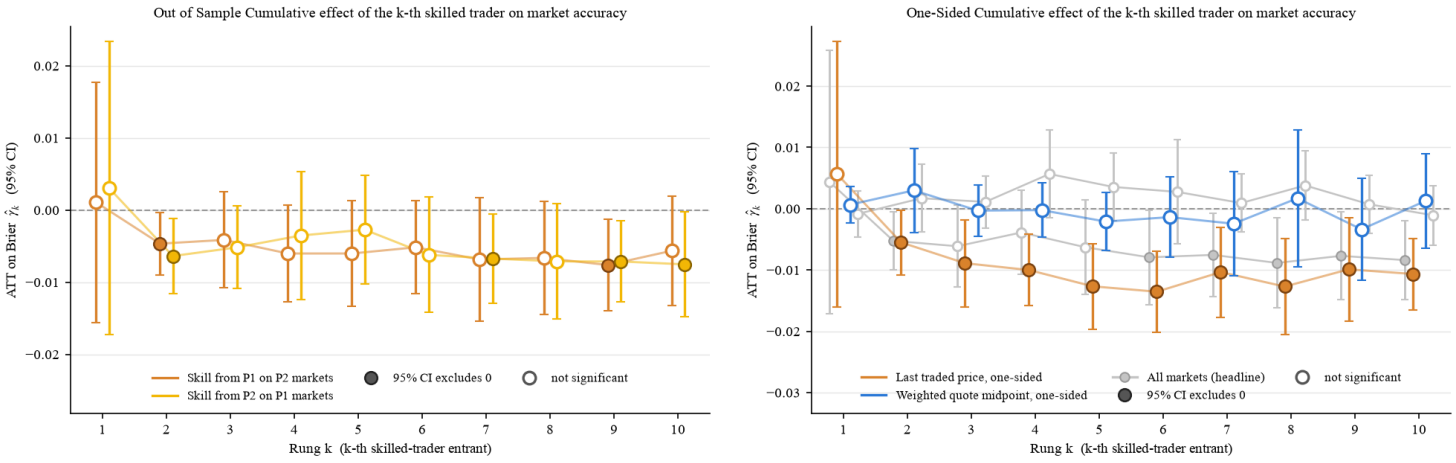


Figure 12: Difference in the time-weighted Brier score to close between markets ending with exactly k skilled traders and those ending with $k-1$.

Out of sample, the pattern holds, with reductions of 0.003 to 0.008 from the second rung onward, significant at four rungs for P2 skill and two for P1 skill (Figure 13(a)). In one-sided markets the cumulative effect is significant at every rung from the second to the tenth and reaches 0.0135 at the sixth. The midpoint estimate is again not distinguishable from zero. All of these estimates depend on how the post-entry window and the volume match are defined.



(a) Out-of-sample skill

(b) One-sided markets

Figure 13: Difference in the time-weighted Brier score to close. Gray markers in (b) repeat Figure 12.