Which forecasts to hold back: the stated probability beats the self-rating

A preregistered study of skipping, hedging and abstain filters in 1,064,355 Good Judgment Project forecasts, with shuffle controls

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F32 in the findings ledger. Strong signal (tier B*).

Abstract

A source that may decline to answer needs a rule for which answers to hold back. We tested two candidate inputs for that rule on the public survey forecasts of the Good Judgment Project: the probability the forecaster stated, and the forecaster's own 1 to 5 rating of their expertise on that forecast. On 410,451 binary forecasts, filters were scored by the area under the risk-coverage curve, the mean Brier score of kept forecasts as coverage falls from 100 to 10 percent. A filter on the stated probability scored 0.171, a filter on the self-rating 0.259, and random order 0.277. Adding the self-rating to the probability gained 0.0026 (95 percent interval 0.0007 to 0.0046), under the preregistered bar of 0.005, and a rating shuffled within each forecaster-year gained 0.0024 on average. We also asked whether people who skip questions forecast better. On first forecasts only, skip rate correlated with better relative Brier score at rho +0.286 (+0.264 to +0.303) across 9,204 forecaster-years, and +0.195 (+0.174 to +0.212) when timing was matched as well. Within the same person across consecutive years the correlation was +0.041 (-0.007 to +0.095) over 1,578 pairs, so no within-person benefit of skipping was shown. Forecasts rated expertise 1 are pulled toward 50 percent by about the right amount: the best extremising factor was 0.92 (0.79 to 1.06).

1 What this means for a reader

When a forecaster, a panel or a software system has to choose which answers to withhold, the number it already states is a much better guide than how expert it says it is. In a large public forecasting tournament, a filter on the stated probability cut the error of kept forecasts about six times as much as a filter on self-rated expertise, and the self-rating added nothing a shuffled rating did not also add. People who answer selectively are better forecasters on average, so a habit of skipping marks a careful source. The data does not show that skipping a question makes the remaining answers better.

2 Introduction and motivation

The Good Judgment Project ran from 2011 to 2015 as one of several research teams in a geopolitical forecasting tournament. Thousands of volunteers gave probabilities for world events, and the project tested training, teaming and tracking of top performers as ways to raise accuracy [2]. The survey forecasts, with each forecaster's self-rated expertise on each forecast for the first three years, are public.

Forecasters in the project chose which questions to answer. That choice is part of the data. Merkle, Steyvers, Mellers and Tetlock [3] asked whether good and poor forecasters choose questions differently, and whether those choices can sharpen estimates of a forecaster's ability beyond what a proper scoring rule reads from the forecasts alone. A proper scoring rule, in decision theory, is a score that a forecaster does best on in expectation by stating their true belief; a rule that pays for shading toward a favoured answer is not proper. Sempere and Lawsen [5] showed that a mean Brier score gives a forecaster a reason to skip. The Brier score, from weather forecasting, is the squared difference between the stated probabilities and what happened, so

lower is better; it scores each answer given and says nothing about questions left unanswered: a forecaster whose average Brier score is b^2 should not forecast any question whose true probability lies between b and $1 - b$, even when they know it exactly. Skipping and scoring are therefore tangled before any data is read.

Machine learning has a matching problem under the name selective prediction. A classifier with a reject option keeps the inputs it is most confident on and declines the rest; the term covers that choice of what to answer and leaves out how the kept answers are made. Its quality is drawn as a risk-coverage curve, the error on kept inputs against the share kept [6]. Here risk is the mean Brier score of kept forecasts and coverage the share kept, so the curve says nothing about the forecasts withheld. We use the same curve to score abstain filters for human forecasts.

The third question concerns the hedge. An earlier round of this study found that a forecast a person rates as expertise 1 is worse than the same person's higher-rated forecasts and sits nearer 50 percent. Extremising, the transformation Baron, Mellers and colleagues studied for aggregated forecasts [4], pushes a probability away from 50 percent by scaling its log-odds by a fixed factor, and a factor below 1 pulls it inward instead; it changes no forecast's direction. It can test whether that pull toward 50 percent is too large, too small or the right size.

The study was built to inform one design choice: how the abstain filter in Flywheel, our accountability software, should rank a source's answers. Should it credit a source for abstaining, and does it need the source's self-rating beside the stated probability?

3 Preregistration and design

This is the second round of the study. The preregistration, a practice from clinical trials and psychology in which a dated plan is fixed before results are seen, was written on 29 October 2026 after the first round's results and before any second-round statistic was computed on the real data, and it was hash-locked at 14:59:09 UTC that day. A plan revised after a look at second-round outcomes would not have counted as one. It lists what had been seen before it was written: the first round's results and structural counts of conditions, first forecasts and repeat forecasters taken from the survey files with no outcome joined.

The second-round code was checked first on synthetic files shaped like the GJP data, with planted effects that the code recovered. That is a planted check: data built so the right answer is known, run through the same analysis before the real data, to show the analysis can find an effect when it is there, which no check on real data can show. That check showed that matching on the number of questions answered removes most of the variation in skip rate, because skip rate is one minus answered over available. Matching on forecast count therefore stayed the primary design, as in the first round. This is recorded in the preregistration.

The file names ten verdicts in four parts, each with rules that must all hold, seven shuffle controls with their draw counts, a prior probability for each verdict, and a decision mapping from outcomes to the filter design. A verdict, in our usage, is the pass or fail of one such rule set read off by code, never a reading written after the numbers were seen. A control is a copy of the data with one link broken, by shuffling, and scored with the same code as the real data; a comparison with another dataset or a published number is not a control. Appendix A gives every verdict and its rule. Nothing was added after the run.

4 Data and provenance

The data are the Good Judgment Project survey forecasts for years 1 to 4 and the question file, published by Mellers and colleagues on the Harvard Dataverse (doi:10.7910/DVN/BPCDH5) under CC0 1.0 [1]. CC0 is the Creative Commons public domain dedication: anyone may reuse the data for any purpose without permission, and citation is a scholarly courtesy rather than a licence term. The person-level test file and the prediction-market files in the same release were not used; in a prediction market people trade contracts that pay on an outcome and the price is read as a probability, which is a different instrument from the survey forecasts used

here. The SHA-256 of each of the five input files was recorded before the run and matched after it. SHA-256 is a cryptographic hash, a 64-character fingerprint that changes if any byte of a file changes; it shows the files were not altered and keeps nothing secret.

The survey files hold 3,143,460 rows, one per answer option. Dropping 43,944 withdrawal rows left 1,234,940 forecasts. Of the 617 questions, 498 closed with an outcome and 119 were voided. A voided question is one the tournament cancelled without scoring, a tournament term with no bearing on the verdict sense of void. We dropped 157,189 forecasts on voided or unresolved questions, 13,381 dated after the question closed and 15 whose probabilities could not be repaired. That leaves 1,064,355 scored forecasts from 9,914 people, of which 414,570 are a person's first forecast on a question in a year.

Self-rated expertise, 1 to 5 on each forecast, exists in years 1 to 3 only; all 619,380 year 4 forecasts lack it. The abstain-filter sample is therefore every binary scored forecast in years 1 to 3 with a rating: 410,451 forecasts. Experimental conditions 1 and 2 are individuals, 4 teams and 5 superteams. A superteam, in the project's design, is a team made up of forecasters drawn from the top performers of the previous year; ordinary teams were formed without that selection. The prediction-market condition is not in the survey files.

The data hold no record of a declined question. A skip is inferred: a closed question that was open during a forecaster's active window in a year and that they never answered. The word is ours for that inferred absence, and it excludes questions that closed before or opened after the forecaster's active window. Only aggregates and distributions are reported here. No score for any named person was stored.

5 Methods

Scoring

Each forecast is scored with the Brier score in the GJP form, summed over the answer options, so that it runs from 0 to 2:

$$B = \sum_{k} (p_k - o_k)^2 \tag{1}$$

where p_k is the stated probability of option k and o_k is 1 if that option occurred and 0 otherwise. The relative score of forecaster i on first forecasts is the mean, over the n_i questions they answered, of their first-forecast Brier score minus the median first-forecast Brier score on that question. It is our measure of skill against the crowd on the same question, and it leaves out how hard the question was, which the median absorbs:

$$S_i = \frac{1}{n_i} \sum_{q} (B_{iq} - m_q) \tag{2}$$

The time-matched version $S_{i,t}$ takes the median m_q only over first forecasts made in the same tenth of the question's open period, drops cells with fewer than 5 first forecasts, and needs 5 or more valid questions per forecaster-year, our unit of analysis: one person's forecasts within one tournament year, so a person active in three years counts three times and no two people are pooled. Lower S is better, so the tests correlate skip rate with $-S$.

Skip rate and the matched correlation

A forecaster-year is eligible with at least 10 available questions and 5 answered. Its skip rate is

$$s_i = 1 - \frac{n_{ans}}{n_{av}} \tag{3}$$

where n_{ans} is the number of available questions answered and n_{av} the number available. Busy forecasters answer more, so the correlation is taken inside strata h of year and twentieth of forecast count, each with 10

or more forecaster-years, and averaged with weights w_h equal to stratum size. A stratum, a term from survey statistics, is a group of units alike on the matched variables, so comparisons inside it cannot be driven by those variables; it does nothing about variables left unmatched:

$$\rho_M = \frac{\sum_h w_h \rho_h}{\sum_h w_h} \tag{4}$$

where ρ_h is the Spearman correlation of s_i with $-S_i$ inside stratum h. Spearman's rho, from rank statistics, is the correlation of the ranks of two quantities, so it reads any steady rise or fall and ignores the shape of the relation and the size of outliers. Intervals come from 1,000 bootstrap draws within strata. Each 95 percent interval is the range holding the middle 95 percent of the draws, and it is no statement that the true value lies inside with 95 percent probability. The bootstrap, a resampling method from statistics, recomputes the number on data resampled with replacement and assumes no textbook distribution.

The within-person test takes pairs of consecutive years in which a person kept the same majority condition. Inside each year pair it computes the Spearman correlation of the change in skip rate with the change in time-matched score, and averages over year pairs weighted by pair count. Changing condition between years would change the platform and the questions, so mixed pairs are left out of the primary test.

The size of the hedge

For a stated probability p the extremised probability is

$$p' = \frac{p^a}{p^a + (1 - p)^a} \tag{5}$$

so the log-odds are multiplied by a. The log-odds of a probability p, a term from statistics, is $\ln(p/(1 - p))$, which is 0 at 50 percent and grows without bound toward certainty; it is a rescaling of the probability and adds no information to it. A value a above 1 means forecasts were too hedged, below 1 too extreme, and 1 right-sized. The factor is fitted by minimum Brier score on a grid from 0.30 to 4.00 in steps of 0.01. Questions were split once at random into two halves; the held-out gain is the Brier score at $a = 1$ minus the Brier score at the a fitted on the other half, summed over both halves and taken per forecast. Held-out, in machine-learning usage, means scored on data the fit never saw, which excludes the in-sample improvement any fitted factor gives. The primary contrast uses the first round's within-person sample: forecaster-years with at least 5 forecasts at expertise 1 and 5 at expertise 2 to 5. Intervals come from 1,000 bootstrap draws over questions.

Abstain filters

A filter ranks forecasts by predicted Brier score and keeps the best-ranked share c, the coverage. Its risk $R(c)$ is the mean raw Brier score of the kept forecasts. Each predictor is an out-of-fold cell mean, computed for each forecast from the other folds only, so no forecast helps predict itself: the self-rating filter uses the five expertise levels, the probability filter ten bins of distance from 0.5 in steps of 0.05, and the combined filter their 50 crossed cells, falling back to the probability bin where a cell has fewer than 30 training forecasts. Ties are broken by one fixed random key. The summary is the area under the risk-coverage curve, taken as the mean risk over 19 coverages. The area is a selective-prediction summary of a whole filter in one number, lower being better, and it excludes coverages below 10 percent:

$$A = \frac{1}{19} \sum_j R(c_j) \tag{6}$$

with j running from 1 to 19 and $c_j = 0.05(j + 1)$, so coverage runs from 0.10 to 1.00. Lower is better. A random filter keeps the overall mean at every coverage.

Controls and p values

Every shuffle control is scored with the same code as the real data. The p value of a comparison, a p value

from draws as in permutation statistics, assumes no textbook distribution and counts the control draws at or above the real value, with one added to the numerator and the denominator:

$$p = \frac{k+1}{n+1} \tag{7}$$

where k is the number of draws at or above the real value and n the number of draws. At 1,000 draws the smallest possible value is 0.001.

6 Controls and planted checks

Seven controls were fixed in the preregistration. Each breaks one link and keeps the rest of the data. The scores a control gives over its draws form a null distribution, whose 95th percentile is the bar a real score must clear; it is drawn from these data, never assumed from theory.

- Skip rate shuffled within year and forecast-count strata, the null for the first-forecast correlation (1,000 draws), and the same on the time-matched table (1,000).
- Change in skip rate shuffled within year pair, the null for the within-person test (1,000).
- The expertise-1 label shuffled within forecaster-year with the question halves fixed, the null for the difference in extremising gain (300).
- Expertise shuffled within forecaster-year with all predictors refitted, the null for the gain from adding the rating to the probability filter (300). This asks whether the rating carries information about one forecast or only about who is rating.
- The team label shuffled across forecaster-years within year, the null for the team-against-individual differences in skip effect and in filter gain (300 each).

The random filter is the known floor: Brier 0.277 at every coverage. Before the real data were read, the code recovered planted effects in synthetic files of the same shape: skippers sharper on first forecasts, a within-person change, and over-hedged expertise-1 forecasts. The first round also carried a positive control that passed, a test of an effect already known to be there, which the instrument must detect for its other readings to be trusted: one point of self-rated expertise improved relative Brier by 0.0226 (0.0185 to 0.0270) within forecaster-years, against a shuffle 95th percentile of 0.0015. Every control in this round had enough spread to be usable.

7 Results

Abstain filters

Figure 1 draws the three risk-coverage curves and Table 1 gives their areas. A filter on the stated probability scored 0.171, a filter on the self-rating 0.259, and random order 0.277. The probability filter gains 0.107 (0.097 to 0.115) over random, and the self-rating filter 0.018 (0.009 to 0.027), about a sixth as much.
Adding the self-rating to the probability gained 0.0026 (95 percent interval 0.0007 to 0.0046), under the preregistered bar of 0.005, and a rating shuffled within each forecaster-year gained 0.0024 on average. The interval excludes zero, so the combined filter is better than the probability alone by a small amount. The shuffle buys nearly all of it: the small edge comes from who rates themselves high on average. The rating given to one forecast adds nothing past that. The combined filter leads only below coverage 0.30. Relative Brier gives the same order: areas 0.124 for self-rating, 0.063 for probability, 0.061 for both and 0.134 for random.

The by-hand rules lose to the probability filter at the same coverage. Withholding every expertise-1 forecast keeps 78 percent of forecasts at Brier 0.267; the probability filter at that coverage reaches 0.214 and the combined filter 0.214. Withholding expertise 1 and 2 keeps 48 percent at 0.259, against 0.141 for the probability filter and 0.142 combined.

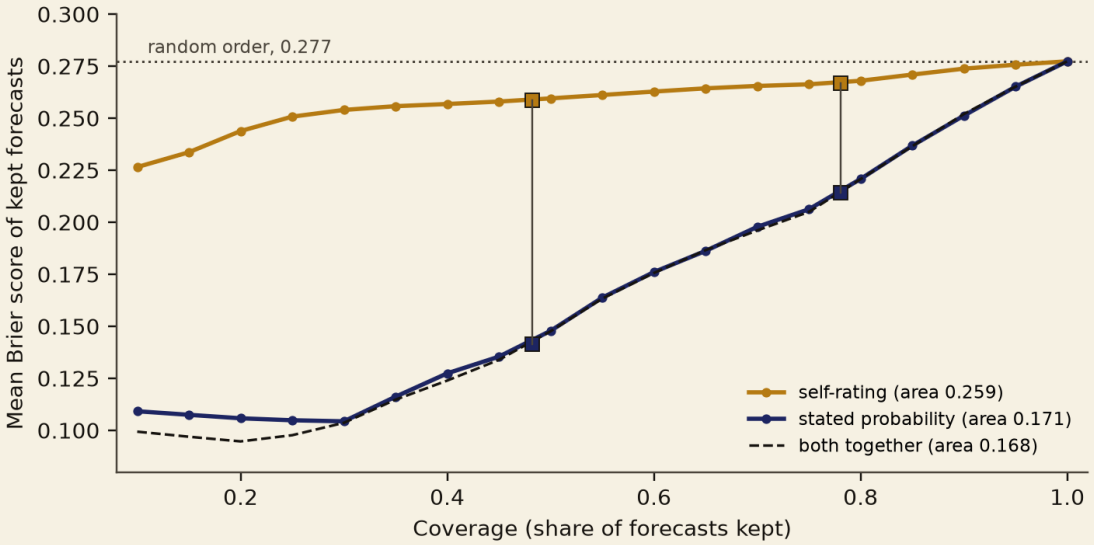


Figure 1. Mean Brier score of kept forecasts against coverage for the three filters, on 410,451 binary forecasts from years 1 to 3. Squares mark two rules a person could apply by hand: withhold every expertise-1 forecast (coverage 0.78) and withhold expertise 1 and 2 (coverage 0.48), each against the probability filter at the same coverage. Drawn from the run's result file.

Table 1. Area under the risk-coverage curve (raw Brier, 0 to 2 scale; lower is better). Intervals are 95 percent bootstraps over questions.

Filter	Area	Gain and interval
Random order	0.277	reference
Self-rated expertise	0.259	0.018 over random (0.009 to 0.027)
Stated probability	0.171	0.107 over random (0.097 to 0.115)
Both together	0.168	0.0026 over probability (0.0007 to 0.0046)
Both, rating shuffled within forecaster-year	not reported	0.0024 over probability on average; 95th percentile 0.0030; p 0.31

Skipping and first forecasts

Figure 2 and Table 2 give the skip correlations. Across the 9,204 eligible forecaster-years (1,781; 1,114; 1,360 and 4,949 in years 1 to 4), the median skip rate was 0.59. On first forecasts only, skip rate correlated with better relative Brier score at rho +0.286 (+0.264 to +0.303) across 9,204 forecaster-years, and +0.195 (+0.174 to +0.212) when timing was matched as well. The first round's statistic on all forecasts, updates included, recomputed on the same table, was +0.351, so removing every update keeps about 80 percent of it.

Skippers enter questions later. Inside strata, skip rate correlates +0.41 with the mean timing of a person's first forecasts, and matching on timing removes about a third of the first-forecast effect. Inside the same strata skip rate also correlates +0.452 with updates per answered question, and updates per question correlate +0.178 with better first-forecast score: people who update a lot also make better first forecasts.

Within the same person across consecutive years the correlation was +0.041 (-0.007 to +0.095) over 1,578 pairs, so no within-person benefit of skipping was shown. The pairs split 501, 469 and 608 across years 1 to 2, 2 to 3 and 3 to 4. The estimate sits just under the shuffle's 95th percentile of +0.041 (p 0.055). The test had room to see an effect of cross-sectional size: test-retest correlation, the psychometric term for how well one person's value in one year predicts the same person's value the next, is 0.62 for skip rate and 0.46 for score, skip rate changed by a standard deviation of 0.25 between years, and the upper bound of +0.095 is under half the time-matched cross-section. All 1,962 pairs, mixed conditions included, give +0.031 (-0.013 to +0.075).

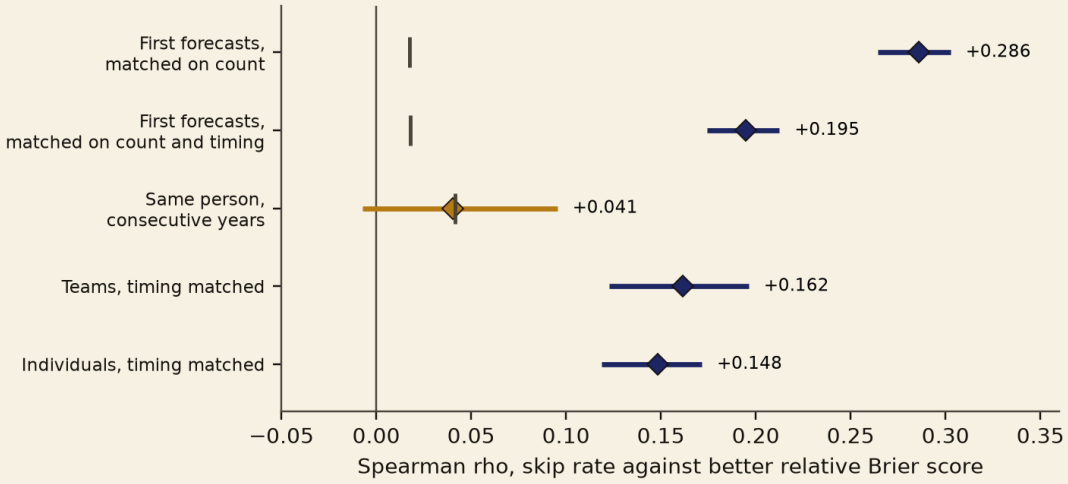


Figure 2. Matched Spearman correlations of skip rate with better relative Brier score, with 95 percent intervals (bars) and the shuffle control's 95th percentile (grey ticks) where the preregistration set one. Drawn from the run's result file.

Table 2. Skip rate against better relative Brier score. Shuffle values are the control's 95th percentile; p from eq. 7.

Test	rho	95 percent interval	Shuffle	p
First forecasts, 9,204 forecaster-years	+0.286	+0.264 to +0.303	+0.018	0.001
First forecasts, timing matched	+0.195	+0.174 to +0.212	+0.018	0.001
Same person, 1,578 pairs	+0.041	-0.007 to +0.095	+0.041	0.055
Teams, timing matched, 3,248	+0.162	+0.123 to +0.196	none	none
Individuals, timing matched, 5,956	+0.148	+0.119 to +0.172	none	none

Teams skip more than individuals (median skip rate 0.63 against 0.57), but the skip effect did not differ detectably: the gap of 0.013 sits under the label-shuffle 95th percentile of 0.043. Superteams removed, condition 4 alone gives +0.154. The gain from adding the self-rating to the filter was 0.0033 in teams and 0.0008 for individuals, a gap of 0.0025 against a shuffle 95th percentile of 0.0030 (p 0.11).

The size of the hedge

In the within-person sample (1,826 forecaster-years; 77,572 forecasts rated 1 and 165,914 rated 2 to 5), forecasts rated expertise 1 are pulled toward 50 percent by about the right amount: the best extremising factor was 0.92 (0.79 to 1.06). The held-out gain from extremising them was +0.0002 (-0.0010 to +0.0016). Forecasts rated 2 to 5 gave 0.93 (0.81 to 1.04), a difference of -0.01 (-0.09 to +0.08). Extremising helped neither group: the gain for low-rated forecasts minus the gain for high-rated ones was -0.0001 (-0.0009 to +0.0008). Table 3 shows the folded calibration behind these numbers.

Table 3. Folded calibration in the within-person sample: the stated probability on the side the forecast leans to, against how often that side happened.

Stated	Rated 1, n	Stated	Hit rate	Rated 2 to 5, n	Stated	Hit rate
0.5 to 0.6	4,664	0.546	0.553	6,968	0.549	0.537
0.6 to 0.7	12,332	0.624	0.618	22,097	0.626	0.602
0.7 to 0.8	12,558	0.727	0.710	25,197	0.728	0.713
0.8 to 0.9	13,360	0.826	0.821	29,612	0.827	0.822
0.9 to 1.0	29,954	0.957	0.911	78,706	0.959	0.930

Folded expected calibration error was 0.024 for low-rated and 0.021 for high-rated forecasts; 4,704 low-rated and 3,334 high-rated forecasts sat at exactly 0.5. Expected calibration error, a forecasting and machine-

learning measure, is the gap between stated probability and hit rate averaged over the bins of Table 3, weighted by bin size; it measures whether forecasts mean what they say and is silent on how informative they are. Across all binary forecasts in years 1 to 3, with people pooled, Figure 3 gives the factor by level. Only the top of the scale misses: forecasts rated 4 or 5 were over-extreme, with a factor of 0.82 (0.68 to 0.96), and at stated probabilities from 0.6 to 0.9 their hit rate ran 0.04 to 0.06 under what they stated.

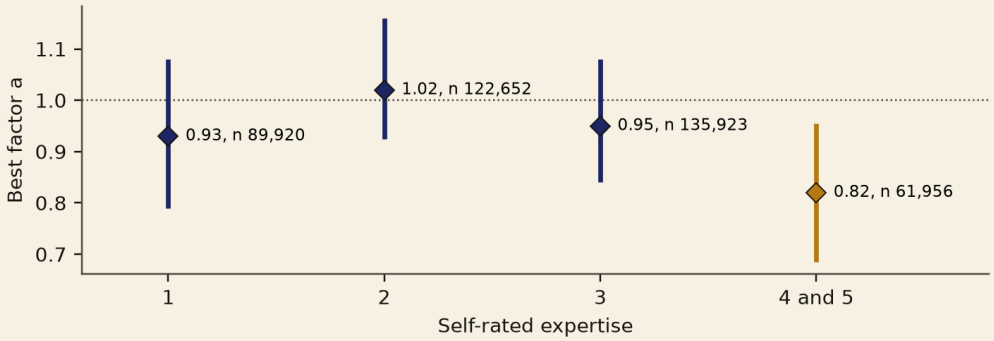


Figure 3. Best extremising factor *a* by self-rated expertise, all binary forecasts in years 1 to 3 with people pooled, with 95 percent intervals over questions. A factor of 1 means right-sized; below 1 means too extreme. Drawn from the run's result file.

8 Discussion

The filter result has a plain mechanism. A forecast near 0 or 1 scores well when it is right, and calibrated forecasters are right about as often as they say. Ranking by stated probability therefore keeps the forecasts that will score best, and part of the gain is arithmetic: the filter rewards confidence by construction. This is the same arithmetic Sempere and Lawsen [5] point to when they show that a mean Brier score pays a forecaster to skip questions in the middle of the probability range. For a filter that only needs to lower the error of what it emits, the arithmetic is the point. For a claim that confident forecasters know more, it is not evidence.

The self-rating adds little because the stated probability already carries it. The first round read "worse and nearer 0.5" as the rating carrying information past the probability. The hedge results settle that point. Expertise-1 forecasts are pulled toward 50 percent by the amount their information warrants, and their worse Brier score is what a calibrated, less informed forecast scores. Once a filter sees the probability, the per-forecast rating adds nothing a within-person shuffle does not also add. The one place the rating says something the probability does not is at the top: forecasters who rate themselves 4 or 5 are over-extreme, so a top rating works as a warning sign for overconfidence.

On skipping, the data separate a trait from an act. People who habitually answer selectively score better on their first forecasts, before any update, at matched timing, in teams and individuals alike. A person who skips more this year than last does not clearly gain from it. Skip rate works as a marker of a careful forecaster; the data do not show that abstaining improves an answer. Merkle and colleagues [3] asked whether good and poor forecasters select questions differently in the same tournament. An earlier search note of ours recorded that they found good forecasters selecting more questions, a sign that runs against ours before matching. We read their abstract only, so we cannot confirm the sign or say whether our forecast-count strata, which compare people of equal activity, account for the difference.

For the filter design, the decision mapping fixed in advance gives a direct reading. The filter ranks by stated probability. It does not take the self-rating as an add-on; a high self-rating is at most a flag to soften a forecast. A habit of abstaining can raise the weight given to a source, but abstention is not credited as something that improves an answer. One setting serves teams and individuals. All of this is human data. For model sources it

sets the design to test, not the setting: a model's stated probabilities need their own check against outcomes.

9 Limitations

- Skips are inferred. A question counted as skipped may have been unseen, or outside the forecaster's condition; "available" is any closed question open during their active window.
- Nothing here is causal. People chose their ratings, their questions and when to forecast.
- The within-person interval reaches +0.095, which leaves room for a small benefit of skipping. Between years the platform, the question mix and the training also changed; same-condition pairs only reduce this.
- Updating, timing and selection remain tangled in the cross-section: skip rate correlates +0.452 with updates per question inside strata.
- A probability filter lowers Brier partly by arithmetic, as the discussion sets out.
- The filters were simulated after the fact on forecasts people chose to make with no filter in place. A live filter would change behaviour.
- Year 4 has no self-ratings, prediction-market trades were not used, and nothing here speaks for language models.
- Merkle and colleagues studied question selection in the same tournament and may disagree on the sign; we read only their abstract.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 every count, correlation, interval, extremising factor, held-out gain, calibration bin, filter area, operating point, shuffle mean, percentile and p value in this paper was re-read from the run's result file and the study's results text; every value the findings ledger also quotes matches it. The ledger is our append-only, dated list of every recorded result; corrections enter it as new lines and old entries are never rewritten. The preregistration's SHA-256 was recomputed from the locked file and matches the hash in the result file and the lock record; the result file's SHA-256 matches the run manifest. All three figures were drawn by a script that reads only the result file. The study was not re-run for this paper. The literature was found by five searches run on 9 November 2026 and read at the level of abstracts, publisher pages and search summaries; the full texts of the cited papers were not read.

11 Corrections

- 29 October 2026: this round narrows the first round's reading. The first round took the low-rating hedge as the self-rating carrying information past the stated probability; this round shows the probability already carries it.
- 9 November 2026: first published version.

12 Reproduction

- Data: download the GJP survey forecasts (years 1 to 4) and the question file from the Harvard Dataverse, doi:10.7910/DVN/BPCDH5.
- Drop withdrawals, forecasts on voided or unresolved questions, forecasts dated after close and unrepairable probabilities. Score each forecast with eq. 1.
- For each forecaster-year with at least 10 available and 5 answered questions, compute the skip rate (eq. 3) and the first-forecast relative score (eq. 2), plain and time-matched. Take the matched correlation (eq. 4) inside year and forecast-count strata, and the within-person version on same-condition consecutive-year

pairs.

- Fit the extremising factor (eq. 5) by grid search with a random split of questions into halves.
- Rank binary forecasts from years 1 to 3 by out-of-fold predicted Brier from self-rating, probability bin and both, and take the area (eq. 6). Shuffle the rating within forecaster-year and refit as the control.
- Seed 2026102902 (numpy generator, consumed in the preregistered order).
- Preregistration SHA-256: be6f2c1a7f6125174f7a03de47baecc080af36002f4a8961ae5077b289084108. Plan SHA-256: 94c658be401adce1a2687fafaab7d1bbae2ff0820a3b4e95d256f011f065ea5f. Result file SHA-256: 7ef5c7a311de888a31ef904f8355a32eb01dd07b2fb9940ab0e88381a129a2e9. Run 20261029T145914Z.
- Cost: 5 minutes 22 seconds of wall time on one desktop computer, controls included.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run, with the prior probability of a pass stated in the same file. A verdict passes only when every rule holds. Values are read from the run's result file.

Table 4. Preregistered verdicts, the rules that decided them, and the prior stated before the run.

Verdict	Deciding rules	Value	Prior	Holds
First forecasts	rho at least 0.05; lower bound above 0; above shuffle 95th percentile (0.018)	+0.286; +0.264	0.6	pass
Timing matched	the same three rules on the time-matched table (shuffle 0.018)	+0.195; +0.174	0.5	pass
Same person	above 0; lower bound above 0; above shuffle 95th percentile (0.041)	+0.041; -0.007	0.3	fail
Hedge right-sized	interval of a for expertise 1 contains 1; held-out gain under 0.002	0.79 to 1.06; 0.0002	0.3	pass
Same correction	interval of the difference in a contains 0	-0.09 to +0.08	0.4	pass
Low-rated gain more	gain difference above 0; lower bound above 0; above shuffle 95th percentile (0.0002)	-0.0001	0.35	fail
Rating beats random	gain over random above 0; lower bound above 0	0.018; 0.009	0.9	pass
Rating adds to probability	gain at least 0.005; lower bound above 0; above shuffle 95th percentile (0.0030)	0.0026; 0.0007	0.35	fail
Teams differ on skipping	gap above label-shuffle 95th percentile (0.043)	0.013	0.35	fail
Teams differ on the filter	gap above label-shuffle 95th percentile (0.0030)	0.0025	0.25	fail

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **proper scoring rule**. A score for probability forecasts on which a forecaster does best in expectation by stating their true belief. Field: decision theory and forecast verification. Excludes scores that pay for shading a forecast toward some answer, and any rule for which questions to answer.
- **Brier score**. The squared difference between the stated probabilities and the outcomes, here summed over the answer options in the GJP form so it runs from 0 to 2; lower is better. Field: forecast verification, from weather forecasting. Excludes questions left unanswered, which it does not score.
- **selective prediction**. Prediction with a reject option, where a system answers only the inputs it ranks most likely to be right and declines the rest. Field: machine learning. Excludes how the kept answers are produced.
- **risk-coverage curve**. The error on kept items against the share kept; here risk is the mean Brier score of kept forecasts and coverage the share kept. Field: machine learning, selective prediction. Excludes any score for

the forecasts withheld.

- **extremising**. Pushing a probability away from 50 percent by multiplying its log-odds by a factor a, with a below 1 pulling it inward instead. Field: forecast aggregation. Excludes changing which side a forecast leans to.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **CC0 1.0**. The Creative Commons public domain dedication, under which anyone may reuse the work for any purpose without permission. Field: copyright and data licensing. Excludes a licence condition requiring citation, which stays a scholarly courtesy.
- **prediction market**. A market in which people trade contracts that pay on an outcome, with the price read as a probability. Field: economics and forecasting. Excludes the individual survey forecasts analysed in this paper.
- **SHA-256**. A hash function that turns a file into a 64-character fingerprint which changes if any byte changes; a recorded hash shows the file was not altered after it was taken. Field: cryptography. Excludes encryption, because the hash keeps nothing secret and cannot be turned back into the file.
- **voided question**. A question the tournament cancelled without scoring, so no forecast on it has an outcome. Field: the Good Judgment Project tournament. Excludes a void verdict in our sense, which concerns a failed planted check or a sample below its minimum.
- **superteam**. A team made up of forecasters drawn from the top performers of the previous tournament year (condition 5). Field: the Good Judgment Project design. Excludes ordinary teams (condition 4), formed without selection on past accuracy.
- **skip**. A closed question that was open during a forecaster's active window in a year and that they never answered, inferred from the data. Field: our usage. Excludes questions outside the active window, and any recorded decision to decline, which the data do not hold.
- **relative Brier score**. A forecaster's mean, over the questions they answered, of their first-forecast Brier score minus the median first-forecast Brier score on that question; lower is better. Field: forecast verification, our usage. Excludes the difficulty of the question, which subtracting the median removes.
- **forecaster-year**. One person's forecasts within one tournament year, the unit for skip rate and relative score. Field: our usage. Excludes pooling of different people, and pooling one person's years.
- **stratum**. A group of units alike on the matched variables, here year and twentieth of forecast count, inside which a statistic is computed before averaging. Field: survey statistics. Excludes control for variables that were not matched.
- **Spearman correlation (rho)**. The correlation between the ranks of two quantities, from -1 to +1, which reads any steady rise or fall. Field: rank statistics. Excludes the shape of the relation and the pull of outliers on the raw values.

- **95% interval**. The range that holds the middle 95 percent of the bootstrap or control draws for a number; we report it beside the number. Field: statistics. Excludes a statement that the true value lies inside with 95 percent probability.
- **bootstrap**. Recomputing a number many times on data resampled with replacement, here within strata or over questions, to see how much it varies. Field: statistics. Excludes an interval from a textbook formula that assumes a distribution.
- **log-odds**. For a probability p, the natural log of p divided by 1 minus p; 0 at 50 percent and unbounded toward certainty. Field: statistics. Excludes any information beyond the probability, of which it is a rescaling.
- **held-out gain**. The drop in Brier score from extremising, with the factor fitted on one half of the questions and scored on the other half. Field: machine learning. Excludes the in-sample improvement any fitted factor gives on the data it was fitted to.
- **out-of-fold prediction**. A prediction for each item made from a fit on the other folds only. Field: machine learning, cross-validation. Excludes any use of the item's own outcome to predict it.
- **area under the risk-coverage curve**. The mean risk over coverages from 10 to 100 percent in steps of 5, one number for a whole filter; lower is better. Field: machine learning, selective prediction. Excludes coverages below 10 percent.
- **p value (from draws)**. The share of control draws scoring at or above the real score, with one added to the count and to the number of draws, so it is never zero. Field: statistics. Excludes a p value from a textbook formula that assumes a distribution.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **positive control**. A test of an effect already known to be present, which the instrument must detect for its other readings to be trusted. Field: experimental design. Excludes a shuffle control, which removes an effect instead.
- **test-retest correlation**. The correlation between one person's value in one year and the same person's value in the next, a measure of how stable the quantity is. Field: psychometrics. Excludes change caused by anything the person did between the two years.
- **expected calibration error**. The gap between stated probability and hit rate, averaged over probability bins and weighted by bin size; folded here by taking each forecast on the side it leans to. Field: forecast verification and machine learning. Excludes how informative forecasts are, since a forecast of the base rate every time can score zero.
- **findings ledger**. The append-only, dated list of every result we have recorded, numbered F1, F2 and on; corrections are added as new dated lines and old entries are never rewritten. Field: our records. Excludes a curated list of successes.

15 References

[1] Mellers, B. and others. Good Judgment Project data: survey forecasts, years 1 to 4, and question file. Harvard Dataverse, CC0 1.0. doi:10.7910/DVN/BPCDH5

[2] Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E. and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. Psychological Science 25(5), 1106-1115. doi:10.1177/0956797614524255

[3] Merkle, E. C., Steyvers, M., Mellers, B. and Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: the questions we choose also matter. International Journal of Forecasting 33(4), 817-832. doi:10.1016/j.ijforecast.2017.04.002

[4] Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E. and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. Decision Analysis 11(2), 133-145. doi:10.1287/deca.2014.0293

[5] Sempere, N. and Lawsen, A. (2021). Alignment problems with current forecasting platforms. arXiv:2106.11248

[6] Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems 30. arXiv:1705.08500