Forecasts people rate low are worse and more hedged, and people who skip more score better

The first round of a preregistered study of self-rated expertise and question skipping in 1,064,355 Good Judgment Project forecasts, with shuffle controls, and what the second round took back

Zain Dana Harper

Version 1.0, 14 November 2026. Finding F28 in the findings ledger, narrowed by F32. Signal (tier B).

Abstract

A source that can say "I am not sure" or decline to answer is only useful if those signals mean something. We tested both on the public survey forecasts of the Good Judgment Project, where volunteers rated their own expertise from 1 to 5 on each forecast and chose which questions to answer. Within a forecaster-year, one point of self-rated expertise went with a relative Brier score better by 0.0226 (95 percent interval 0.0185 to 0.0270) across 3,780 forecaster-years. Forecasts rated expertise 1 scored 0.0218 worse (0.0155 to 0.0285) than the same person's forecasts rated 2 to 5 and sat 0.055 nearer to 50 percent (0.051 to 0.058), across 1,826 forecaster-years. The preregistered prediction was that low-rated forecasts would be worse and no less hedged, and it failed on the hedging rule. With question-by-time medians the gap fell to 0.0156 (0.0094 to 0.0219), under the preregistered bar of 0.02. At matched forecast count, skip rate correlated with better relative Brier score at rho +0.351 (+0.331 to +0.368) across 9,204 forecaster-years. Inside the same strata skip rate correlated +0.452 with updates per answered question, so this round does not separate skipping from updating. A second round, published as F32, removed the updates and narrowed two of these readings; this paper states what it changed.

1 What this means for a reader

When people in a large forecasting tournament said they knew little about a question, their forecast was a little worse, and they had already moved it toward 50 percent. People who answered fewer of the questions open to them scored better on the ones they did answer. Both patterns are associations in people who chose their own ratings and questions. The follow-up study found that the stated probability already carries what the self-rating knows, and that a person who skips more in one year than the last does not clearly score better, so a habit of skipping marks a careful forecaster and is not shown to improve an answer.

2 Introduction

The Good Judgment Project ran from 2011 to 2015 inside a geopolitical forecasting tournament. Thousands of volunteers gave probabilities for world events, and the project tested training, teaming and tracking as ways to raise accuracy [2]. Its survey forecasts are public, and for the first three years each forecast carries the forecaster's own 1 to 5 rating of their expertise on that question.

Two things in those files bear on a practical design question: how software that collects answers from many sources should treat a source's stated confidence, and whether it should let a source decline. The first is the self-rating. A review of ways to identify skilled forecasters lists the same 1 to 5 item among expertise-based predictors and attributes it to the project's own papers [5]. The second is the choice of questions. Merkle, Steyvers, Mellers and Tetlock [3] pointed out that a Brier score, the squared difference between stated probabilities and outcomes, looks only at the forecasts a person submitted, and asked whether the questions

a person selects add information about their ability. A Brier score is lower when better, and it gives no score to a question left unanswered.

We asked two questions of the data. Within one person, does a low self-rating mark a worse forecast, and is that forecast already hedged? Across people, do those who skip more score better on what they answer? The decision the study informed was whether an abstain filter in Flywheel, our accountability software, can use a source's own confidence report or needs an outside measure. Our searches for this paper returned the two works above and the project's own account of its best forecasters [4]; none of the search results gave an effect size for the self-rating inside one person, which is what a search returned and no proof that none is published.

3 Preregistration and design

The preregistration, a dated plan fixed and hash-locked before any result is seen, was written on 29 October 2026 and locked at 14:29:13 UTC that day. Before writing it we had seen file lists, column headers, and counts of questions, forecasts and rating levels with no outcome joined to any forecast. No Brier score had been computed on the real data. The code had been run on synthetic files shaped like the real ones, with effects built in that it recovered, which is a planted check: data with a known answer run through the same analysis first, to show the analysis can see an effect when one is there.

The plan names seven verdicts, each a set of rules that must all hold, read off by code. A verdict in our usage is that pass or fail and never a reading written after the numbers were seen. It also states a prior probability for each main verdict (0.8, 0.25 and 0.35 for the positive control, the low-rating prediction and the skipping prediction) and a mapping from outcomes to the filter design. Appendix A gives every rule.

4 Data and provenance

The data are the Good Judgment Project survey forecasts for years 1 to 4 and the question file, published on the Harvard Dataverse under CC0 1.0, the Creative Commons public domain dedication, which allows any reuse and leaves citation as a courtesy [1]. The SHA-256 of each of the five input files was recorded in the run's result file.

The survey files hold 3,143,460 rows, one per answer option. Dropping 43,944 withdrawal rows left 1,234,940 forecasts. We dropped 157,189 forecasts on questions that were voided or had no recorded outcome, 13,381 dated after the question closed and 15 whose probabilities could not be repaired. That leaves 1,064,355 scored forecasts. Self-rated expertise exists in years 1 to 3 only, on 410,451 scored binary forecasts. Only aggregates and distributions were stored; no value for any single forecaster was written to any output.

The files 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 excludes any recorded decision to decline, which the data do not have, so a skipped question may have been unseen.

5 Methods

Scores

Each forecast is scored with the Brier score in the project's form, summed over the answer options, so 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 Brier score of a forecast is B minus the median B of all scored survey forecasts on the same question, so it

measures skill against the crowd and leaves out how hard the question was. The secondary form takes the median inside the question and the fifth of its open period in which the forecast was made. Extremity is the distance of a binary forecast from 50 percent, $|p - 0.5|$, from 0 to 0.5; it measures how hedged a forecast is and says nothing about whether it was right. The unit is the forecaster-year, one person's forecasts within one tournament year, so a person active in three years counts three times.

The within-person tests

The positive control tests an effect that should be there for the other readings to be trusted: that self-rated expertise goes with accuracy at all. For forecaster-years with at least 10 scored binary forecasts using two or more rating levels, relative Brier score and rating were both centred within the forecaster-year, and one slope β was fitted to the pooled data. The reported quantity is $-\beta$, so a positive value means a higher rating goes with a better score.

The low-rating test uses forecaster-years with at least 5 forecasts rated 1 and 5 rated 2 to 5. For each it takes

$$d_B = S_{low} - S_{rest} \tag{2}$$

where S_{low} is the mean relative Brier score of the forecasts rated 1 and S_{rest} that of the forecasts rated 2 to 5, and the matching difference in extremity, d_E. The statistic is the mean over forecaster-years. The preregistered bar for d_B was 0.02 on the 0 to 2 scale.

Skip rate and the matched correlation

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

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

with n_{ans} the available questions answered and n_{av} the number available. Busy forecasters answer more, so the correlation is taken inside each stratum h of year and twentieth of forecast count, a group alike on the matched variables, and averaged with weights w_h equal to stratum size:

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

where ρ_h is the Spearman rho, the correlation of ranks, between skip rate and the negative of the forecaster-year's mean relative score. Strata with fewer than 10 forecaster-years were dropped. Matching controls only the matched variables.

Intervals

Every 95 percent interval is the middle 95 percent of bootstrap draws, a resampling of forecaster-years with replacement that assumes no textbook distribution: 2,000 draws for the low-rating test and 1,000 for the others, the skip test resampling inside strata.

6 Controls and planted checks

Each control is a copy of the data with one link broken by shuffling and scored with the same code; a comparison with another dataset is not a control. Three were fixed in advance, with 1,000 draws each, and the 95th percentile of each one's null distribution is the bar the real value must clear.

- Ratings shuffled within forecaster-year, the null for the mean of d_B.
- Ratings shuffled within forecaster-year, the null for the positive-control slope.
- Skip rate shuffled within year and forecast-count strata, the null for ρ_M.

The activity check asks whether matching did its job: the mean correlation of skip rate with forecast count inside strata must be under 0.2 in size. A p value from draws is the share of control draws at or above the real

value, with one added to the count and to the number of draws, so 1,000 draws cannot give less than 0.001.

7 Results

Self-rating inside one person

The positive control passed. Within a forecaster-year, one point of self-rated expertise went with a relative Brier score better by 0.0226 (95 percent interval 0.0185 to 0.0270) across 3,780 forecaster-years. The shuffle's 95th percentile was 0.0015 (p 0.001), and the test covered 398,463 forecasts. Each point of rating also went with a forecast 0.037 further from 50 percent (0.034 to 0.040).

Forecasts rated expertise 1 scored 0.0218 worse (0.0155 to 0.0285) than the same person's forecasts rated 2 to 5 and sat 0.055 nearer to 50 percent (0.051 to 0.058), across 1,826 forecaster-years. The sample held 243,486 forecasts, 77,572 of them rated 1. The score gap cleared the shuffle's 95th percentile of 0.0048 (p 0.001) and cleared the 0.02 bar by 0.0018. The verdict still failed, because the prediction was "worse and no less hedged" and the extremity rule required the lower bound of d_E to be above -0.02; it was -0.058. The effect is small beside the spread between people: the low-rated forecasts were the worse ones in 55 percent of forecaster-years, and the median gap was 0.018.

With question-by-time medians the gap fell to 0.0156 (0.0094 to 0.0219), under the preregistered bar of 0.02. Figure 1 and Table 1 give the three within-person effects.

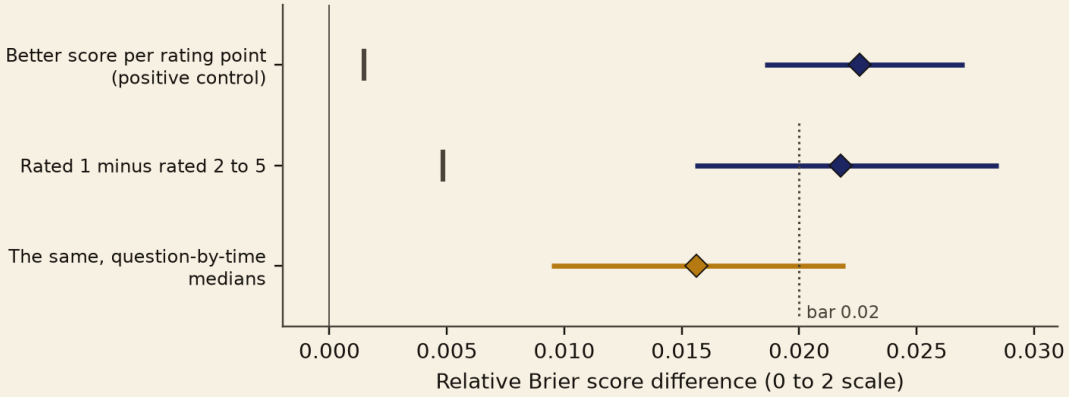


Figure 1. Within-person effects on relative Brier score (0 to 2 scale), with 95 percent intervals (bars), the shuffle control's 95th percentile where one was set (grey ticks) and the preregistered bar of 0.02 for the two low-rating gaps (dotted line). Drawn from the run's result file.

Table 1. Within-person effects. Shuffle values are the control's 95th percentile.

Effect	Value	95 percent interval	Shuffle	p
Better score per rating point, 3,780 forecaster-years	0.0226	0.0185 to 0.0270	0.0015	0.001
Rated 1 minus rated 2 to 5, score, 1,826	+0.0218	+0.0155 to +0.0285	0.0048	0.001
The same with question-by-time medians	+0.0156	+0.0094 to +0.0219	none	none
Rated 1 minus rated 2 to 5, extremity	-0.0548	-0.0584 to -0.0514	none	none

Skipping

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. At matched forecast count, skip rate correlated with better relative Brier score at rho +0.351 (+0.331 to +0.368) across 9,204 forecaster-years. The shuffle's 95th percentile was +0.017 (p 0.001). The activity check

passed: skip rate correlated -0.613 with forecast count overall and -0.078 inside strata. Matching matters to the size of the number. Unmatched, the correlation was +0.119. Matched on questions answered in place of forecasts it was +0.236 (+0.216 to +0.252), and with question-by-time medians +0.211 (+0.188 to +0.229). Figure 2 draws the four.

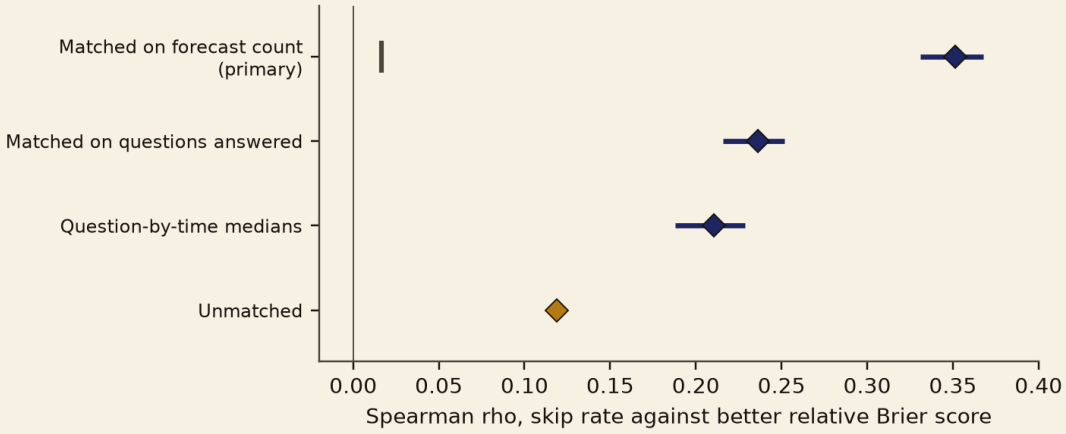


Figure 2. Spearman correlations of skip rate with better relative Brier score across 9,204 forecaster-years, with 95 percent intervals where the plan set one and the shuffle control's 95th percentile for the primary statistic (grey tick). Drawn from the run's result file.

Inside the same strata skip rate correlated +0.452 with updates per answered question, so this round does not separate skipping from updating. The preregistration named this before the run: at a fixed forecast count, a person who answers fewer questions must update each one more, and updating goes with accuracy in this tournament.

Table 2 sets the priors against the outcomes.

Table 2. Priors stated before the run against the outcomes.

Verdict	Prior of a pass	Outcome
Positive control	0.8	pass
Low-rated forecasts worse and no less hedged	0.25	fail: worse by 0.0218, and 0.055 more hedged
Skippers score better at matched activity	0.35	pass, with the updating confound the plan warned of

8 Discussion

The self-rating means something inside one person: a forecast rated low is a worse forecast. People had also acted on that knowledge already, since the low-rated forecasts sit nearer 50 percent. The decision mapping written before the run read this outcome as "self-report modulates the stated probability and is usable as one input" to a filter. The first-round write-up went one step further and read "worse and hedged" as the rating carrying information past the stated probability.

The second round, F32, tested that step and took it back. It found that forecasts rated 1 are pulled toward 50 percent by about the right amount, with a best rescaling factor of 0.92 (0.79 to 1.06) where 1 means right-sized, and that adding the self-rating to a filter on the stated probability gained 0.0026 in area under the risk-coverage curve against a bar of 0.005, while a rating shuffled within each forecaster-year gained 0.0024. A calibrated forecast from a less informed person scores worse, and its stated probability already says by how much.

F32 also removed the confound on skipping. Counting each person's first forecast on a question only, the matched correlation was +0.286 (+0.264 to +0.303), and +0.195 with timing matched as well, so most of the

+0.351 reported here survives the removal of every update. Within the same person across consecutive years it was +0.041 (-0.007 to +0.095) over 1,578 pairs, which shows no benefit to a person from skipping more. The supported reading is a trait: people who answer selectively tend to be better forecasters.

Merkle and colleagues [3] studied question selection in the same tournament. We read their abstract only, in a search snippet that was cut short, so we do not compare signs or sizes with their result.

9 Limitations

- Nothing here is causal. People chose their ratings, their questions and when to forecast.
- Skips are inferred. A question counted as skipped may have been unseen, and "available" means any closed question open during the forecaster's active window.
- The skip correlation in this round includes updates. F32 puts the first-forecast value at +0.286 and found no within-person effect.
- The low-rating gap clears its 0.02 bar by 0.0018, with an interval from 0.0155 to 0.0285, and falls under the bar with question-by-time medians. The 0.02 bar was a choice.
- The reading that the self-rating adds information past the stated probability did not survive F32.
- Year 4 has no self-ratings, prediction-market trades were not used, and nothing here speaks for language models or for any named person.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 14 November 2026 every count, slope, gap, correlation, interval, shuffle percentile, p value and verdict rule in this paper was re-read from the run's result file. They match the findings ledger, our append-only dated list of recorded results. The preregistration's SHA-256 was recomputed from the locked file and matches the hash in the result file and the lock record. The priors were read from the preregistration. The F32 numbers quoted in the discussion were re-read from the published F32 paper. Both 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 two searches run on 14 November 2026 and read at the level of search snippets; no abstract or full text was opened.

11 Corrections

- 29 October 2026: the second round (F32) narrowed two readings of this round. The skip correlation of +0.351 includes updates; on first forecasts it is +0.286, and no within-person benefit of skipping was shown. The self-rating was read here as carrying information past the stated probability; F32 found the probability already carries it.
- 14 November 2026: narrowed the published claim to an association across and within people because the ledger's first reading, that self-report is usable as an input with signal past the stated probability, did not survive F32.
- 14 November 2026: first published version.

12 Reproduction

- Data: Good Judgment Project survey forecasts, years 1 to 4, and the question file, Harvard Dataverse, CC0 1.0, 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 and subtract the question's median score.
- Fit the pooled within-forecaster-year slope of relative score on rating. Compute eq. 2 and the extremity

difference for forecaster-years with 5 or more forecasts in each arm.

- For each forecaster-year with at least 10 available and 5 answered questions, compute eq. 3 and take eq. 4 inside year and forecast-count strata.
- Shuffle ratings within forecaster-year and skip rates within strata, 1,000 times each, and rescore.
- Seed 20261029 (numpy generator, consumed in the preregistered order).
- Preregistration SHA-256: 723e0f541d7fe7ade90a45e04e6c3c15e853955f1a903c8ff5f75f2f952d12cb. Plan SHA-256: 71bedc68ab6171ff3526e5d0eaa0502204ee4891e1a58103345152786b7bb720. Run 20261029T142954Z.
- Cost: about 154 seconds for the analysis and 107 seconds for the controls on one desktop computer.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the run's result file.

Table 3. Preregistered verdicts and the rules that decided each one.

Verdict	Deciding rules	Value	Holds
Positive control	slope above 0; lower bound above 0; above shuffle 95th percentile (0.0015)	0.0226; 0.0185	pass
Low rating: worse, no less hedged	gap at least 0.02; lower bound above 0; above shuffle 95th percentile (0.0048); extremity lower bound above -0.02	0.0218; 0.0155; -0.058	fail
The same, question-by-time	gap at least 0.02; lower bound above 0; extremity lower bound above -0.02	0.0156; 0.0094; -0.058	fail
Skippers score better	rho at least 0.05; lower bound above 0; above shuffle 95th percentile (0.017)	+0.351; +0.331	pass
Activity check	skip against forecast count inside strata under 0.2 in size	0.078	pass
Matched on questions answered	rho above 0; lower bound above 0	+0.236; +0.216	pass
Question-by-time	rho above 0; lower bound above 0	+0.211; +0.188	pass

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.

- **Brier score**. The squared difference between the stated probabilities and the outcomes, here summed over the answer options in the Good Judgment Project 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.
- **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 (meta-science). 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.

- **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.
- **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 forecast's Brier score minus the median Brier score of all scored survey forecasts on the same question; lower is better. Field: forecast verification, our usage. Excludes the difficulty of the question, which subtracting the median removes.
- **extremity**. The distance of a binary forecast from 50 percent, from 0 for a coin-flip forecast to 0.5 for a certain one. Field: judgment and forecasting research. Excludes whether the forecast was right, and how well calibrated the forecaster is.
- **forecaster-year**. One person's forecasts within one tournament year, the unit for every statistic here. Field: our usage. Excludes pooling of different people, and pooling one person's years.
- **positive control**. A test of an effect expected 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.
- **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 rho**. The correlation between the ranks of two quantities, from -1 to 1, which counts any steadily rising or falling relation and is not pulled by single extreme values. Field: statistics. Excludes the slope of a relation, and evidence of cause.
- **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 statistic many times on forecaster-years resampled with replacement to read how much the estimate varies. Field: statistics. Excludes an interval from a textbook formula that assumes a distribution.
- **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.
- **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.
- **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.
- **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., Mur-

ray, 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] Mellers, B. and others (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science 10(3), 267-281.

[5] Atanasov, P. and others (2023). Talent spotting in crowd prediction. Cited from a search result; the venue was not confirmed.