Proxy Scores as a Real Time Forecaster Evaluation Tool

Mark Himmelstein, David Budescu & Emily Ho

This work was supported by an NSF (DRMS) grant #1919055

Forecaster Evaluation

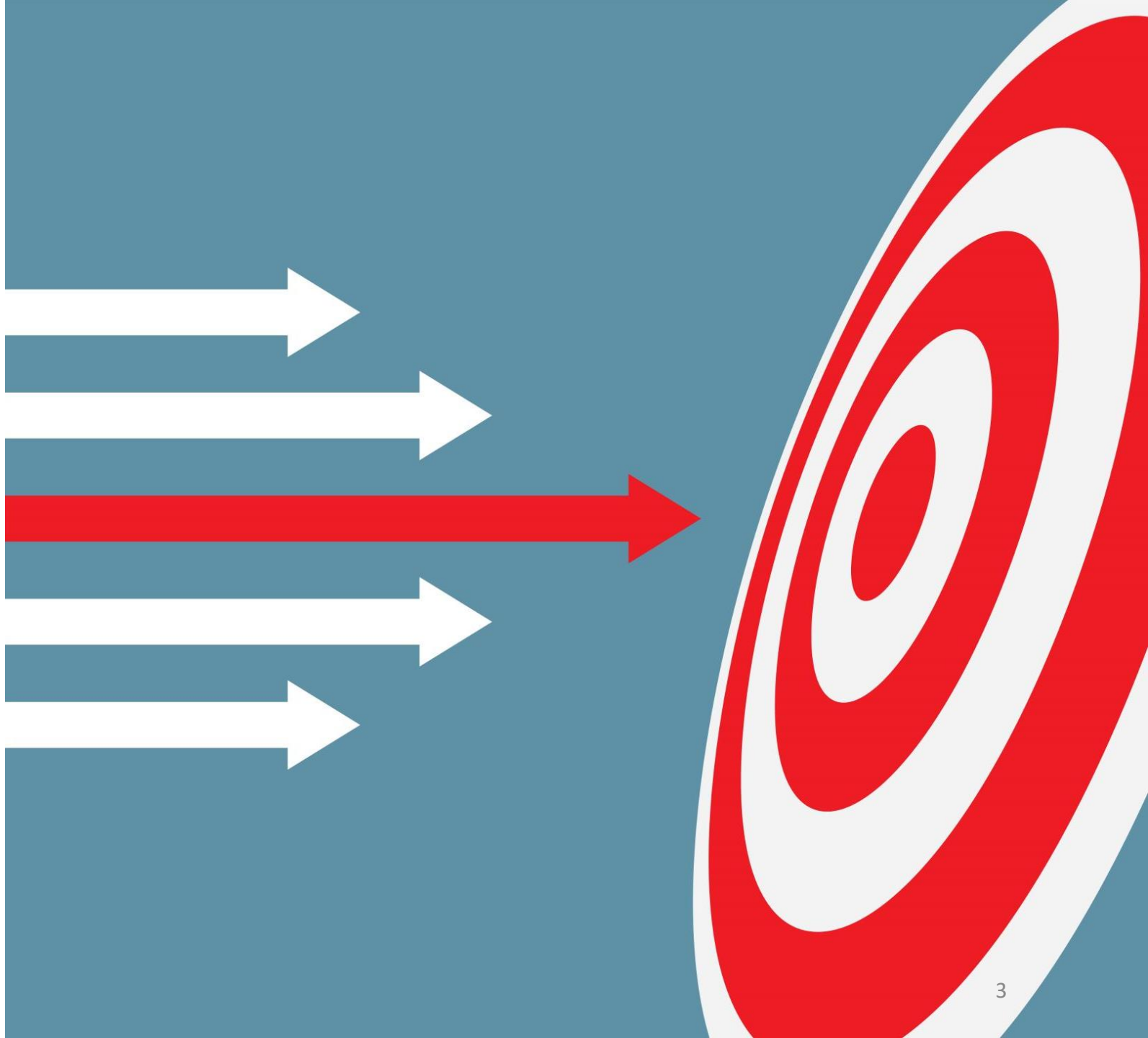
Numerous studies have established that probabilistic forecasting is a consistent and stable skill

Forecasters' accuracy can be predicted by several factors, including:

- General intelligence/numerical reasoning ability (Himmelstein et al., 2021; Mellers et al., 2015)
- Probabilistic calibration (Aspinall, 2010) and coherence (Ho, 2020)
- Tendency to incrementally revise one's beliefs (Atanasov et al., 2017)

But there is one factor that consistently predicts accuracy better than any other

Past Accuracy!



Reliability of Accuracy

Knowing something about how accurate a forecaster has been on average can help predict how accurate they are likely to be moving forward.

Problem: To assess the accuracy of a forecast, the "ground truth" must be known.

Forecasters' average accuracy cannot be assessed until they have forecasted at least some events that have already **resolved.**

Also known as **Cold Start Problem.**

Wisdom of Crowds as Proxy for Truth

Averaging the beliefs of many forecasters tends to produce forecasts that are more accurate than individual members of the crowd, aka the **Wisdom of Crowds** (Budescu & Chen, 2015; Surowiecki, 2005).

Studies have proposed using the crowd's forecast as a **proxy** for the ground truth, to assess accuracy before outcomes are known (Liu et al., 2020; Witkowsi et al., 2017)

Research Questions

- How well do Proxy Scores correlate with actual accuracy?
- Do Proxy Scores from one set of questions predict actual accuracy on other questions?
- Can Proxy Scores identify high performing forecasters without knowing anything about their actual accuracy?
- Can Proxy Scores help us improve on current Wisdom of Crowds methods?

Scoring Rules

Probabilistic accuracy is often assessed using **proper scoring rules**, such as the Brier score

$$BS = \sum_{b=1}^{K} (f_b - o_b)^2$$

K = The number of possible outcomes associated with item
f = forecast value
o = outcome (0 = did not occur, 1 = did occur)

Brier Scores range from 0 (perfect accuracy) to 2 (absolute inaccuracy)

Proxy Score Candidates

Distance Score (DS)

$$DS = \sum_{b=1}^{K} (f_b - c_b)^2$$

c = crowd forecast (aggregate)

Expected Brier Score (EBS)

$$EBS = \sum_{b=1}^{K} c_b \sum_{t_b=1}^{K} (f_{t_b} - o_{t_b})^2$$

When $b = t_b$, $o_{t_b} = 1$, otherwise $o_{t_b} = 0$

EBS is the average of the Brier Scores for all possible outcomes, weighted by the probability the crowd assigns to that outcome.

The measures are highly correlated, but not identical. Most results generalize to both

Today's results focus on EBS

EBS Example:

When will I finish my Dissertation

Before the end of 2022: .15

After end of 2022: .85

Before the end of 2022: .25

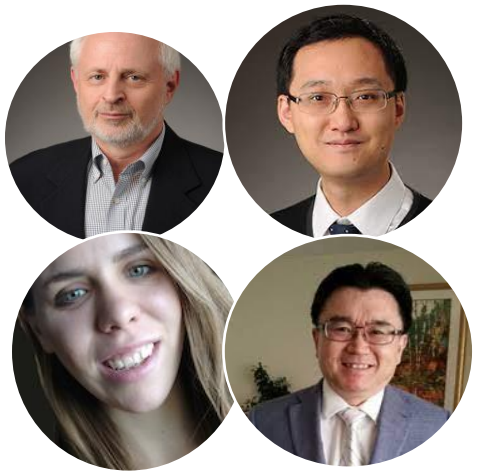
After end of 2022: .75

Possible Brier Scores

Before the end of 2022:
$(.15 - 1)^2 + (.85 - 0)^2 = 1.445$

After the end of 2022:
$(.15 - 0)^2 + (.85 - 1)^2 = .045$

Let's ask a crowd of experts!



My Expected Brier Score

.25(1.445)
+
.75(0.045)
= **0.395**

Design of Study

- Incentivized longitudinal study consisting of 5 waves (three weeks apart) in which the judges forecasted the same events
- Judges forecasted the target events and rated their confidence
- 9 items used a common (remote) resolution horizon
- 3 items used a short resolution horizon (before the next round)

Sample and Incentives

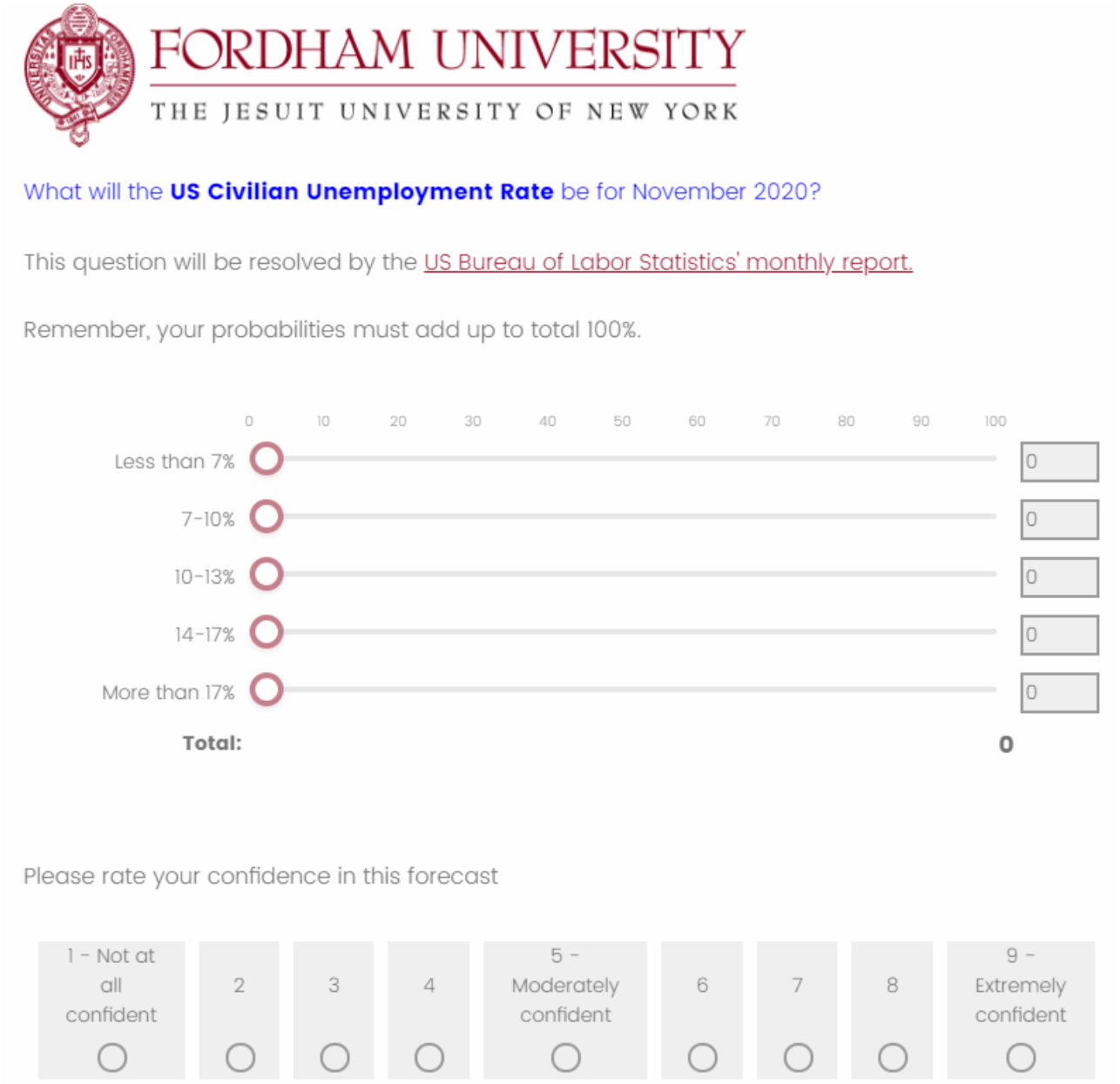
- Total N = 406
  - (Cloud Research; 100 HITs; 95% approval)
  - Mean age = 22.6, SD = 12.2
  - 54% male; 45% female, 1% other
  - 11% HS; 32% some college; 43% bachelor's and 14% graduate degree
- Judges who participated in all waves N = 175
  - Mean age = 23.2, SD = 12.3
  - 59% male; 40% female; 1% other
  - 14% HS; 34% some college; 37% bachelor's and 15% graduate degree
  - **Most analyses focus on this sample**
- Incentives
  - $6 for initial participation (includes intake)
  - $3 for each additional wave
  - Entry into a lottery for one of five $20 bonuses for each wave
  - $15 bonus for most accurate forecaster from each wave
  - $15 bonus for five most accurate forecasters overall (who participated in at least 3 waves)
  - $5 bonus for having "suggested question" selected

Wave	Started on Date	Number of Judges Forecasting	
		Total	Original
1	08/11/2021	302	302
2	09/1/2021	309	258
3	09/22/2021	316	248
4	10/13/2021	311	215
5	10/27/2021	304	222

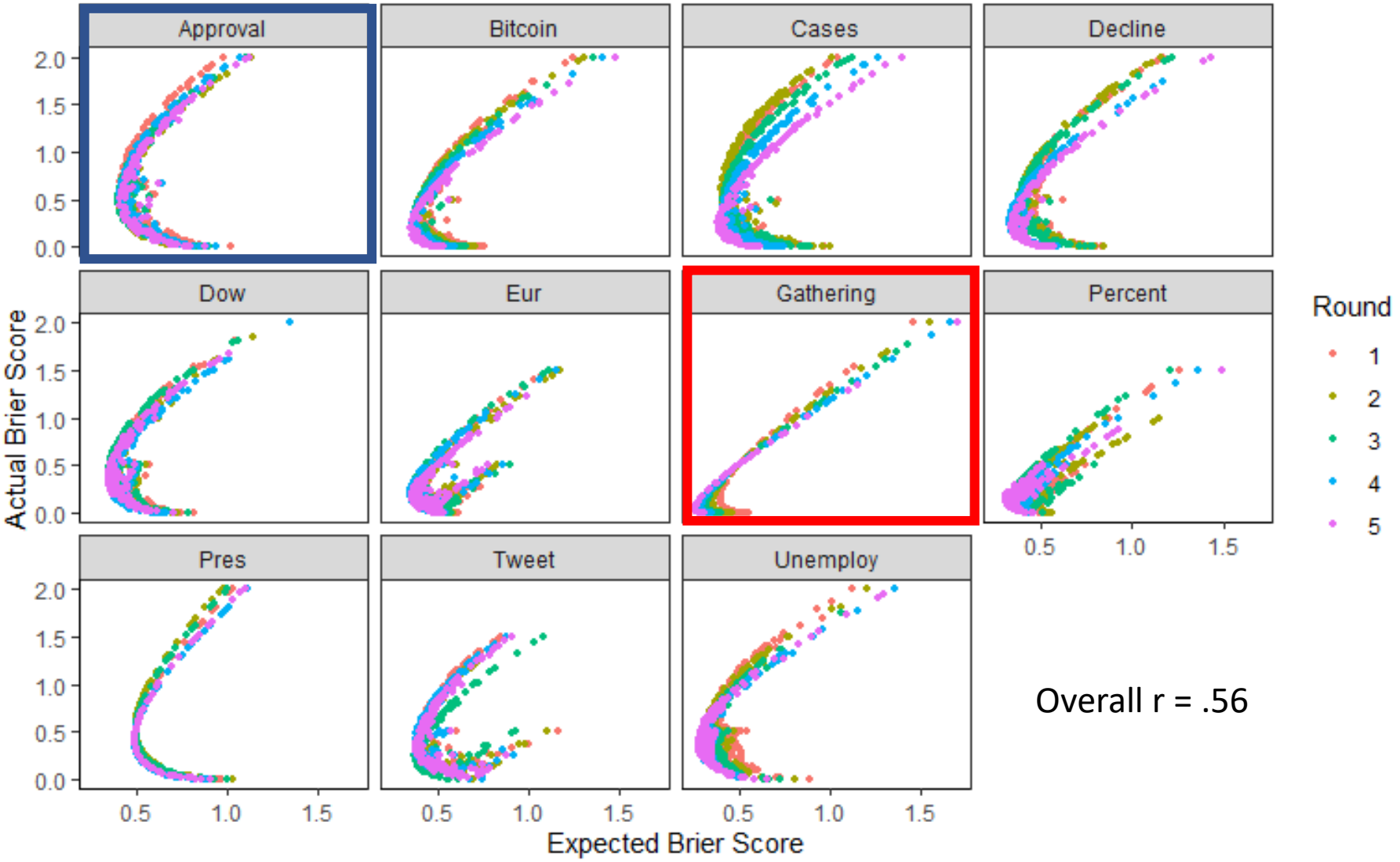
List of Events (with K bins)

Resolution Date	Economics	Politics	COVID
Fixed	What will the Dow Jones index be at the close of the US market on November 4th, 2020? **K = 5**	What will Donald Trump's approval rating be on November 4th, 2020? **K = 4**	How many new cases of COVID-19 will be confirmed on November 4th, 2020 in the United States? **K = 5**
	What will the price of one Bitcoin be (in US dollars) at the end of the day on November 4th, 2020, 11:59 PM Eastern Standard Time? **K = 5**	Who will win the 2020 United States Presidential Election? **K = 2**	How many US States will have fewer confirmed COVID-19 cases between November 11th and 16th than they did between November 4th and November 10th or 2020? **K = 5**
	What will the US Civilian Unemployment Rate be for November 2020? **K = 5**	Which party will hold the most seats in the Senate following November elections, to be sworn in January 2021? **(K = 2)** EXCLUDED!	On November 4th, 2020, will all states permit gatherings of 500 or more people? **K = 2**
Variable	What will the exchange rate of Euros to one (1) U.S. dollar be at the end of the day on [date]? **K = 5**	How many total tweets will be posted by Donald Trump on [date]? **K = 5**	How many U.S. States will have a weekly decline in percentage of positive cases on [date]? **K = 5**

Ordinal Forecasting Question Example



Correlations between BS and EBS for each Forecast



Overall r = .56

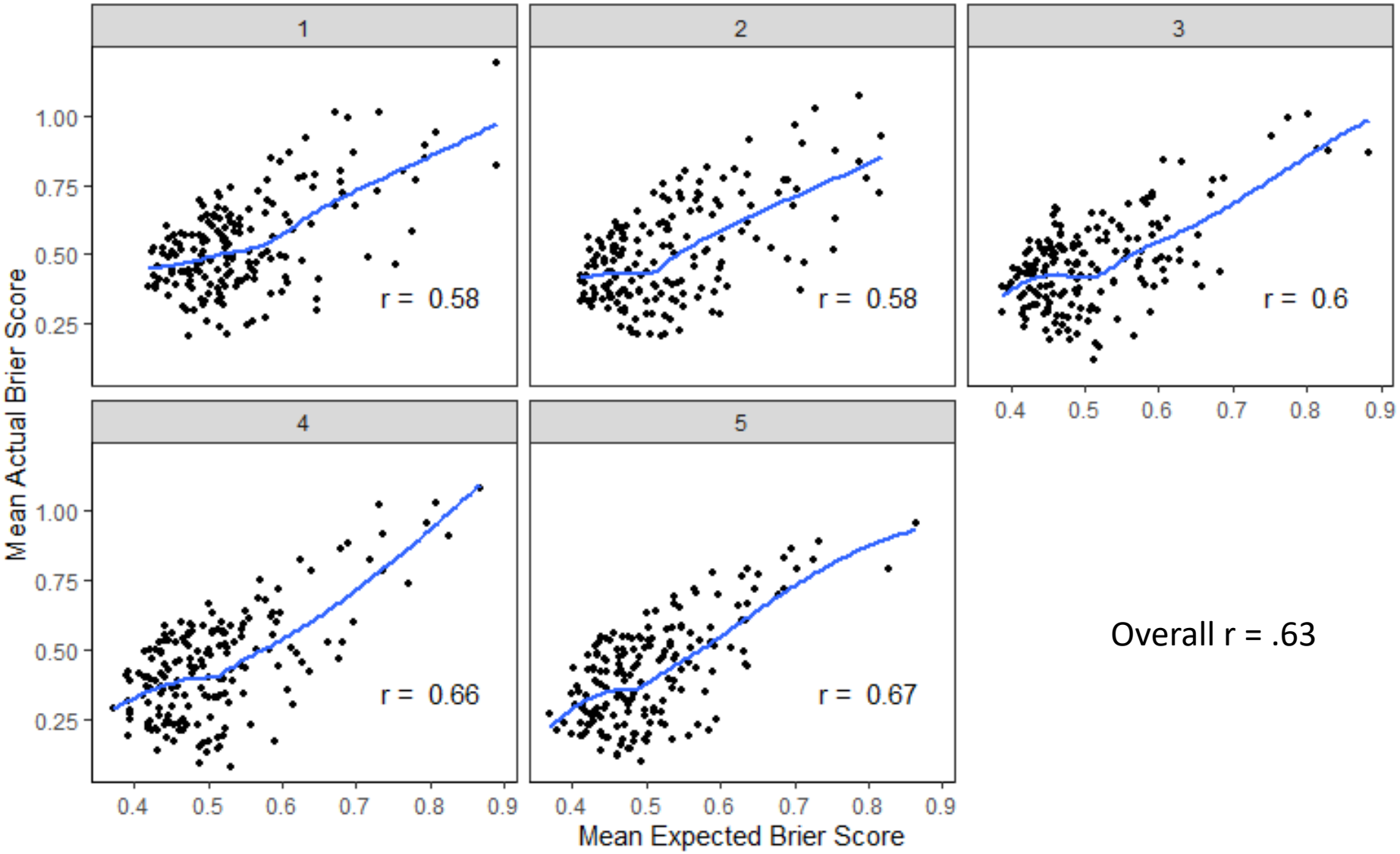
Curved pattern reflects "confident" forecasts receiving very low Brier scores

BS is minimized by P = 1 on "correct" option

EBS is minimized by proximity to crowd forecast

E.g., Crowd forecast was "flat" for Approval, but "extreme" for Gathering

Correlations between forecasters' Mean BS and Mean EBS in Each round

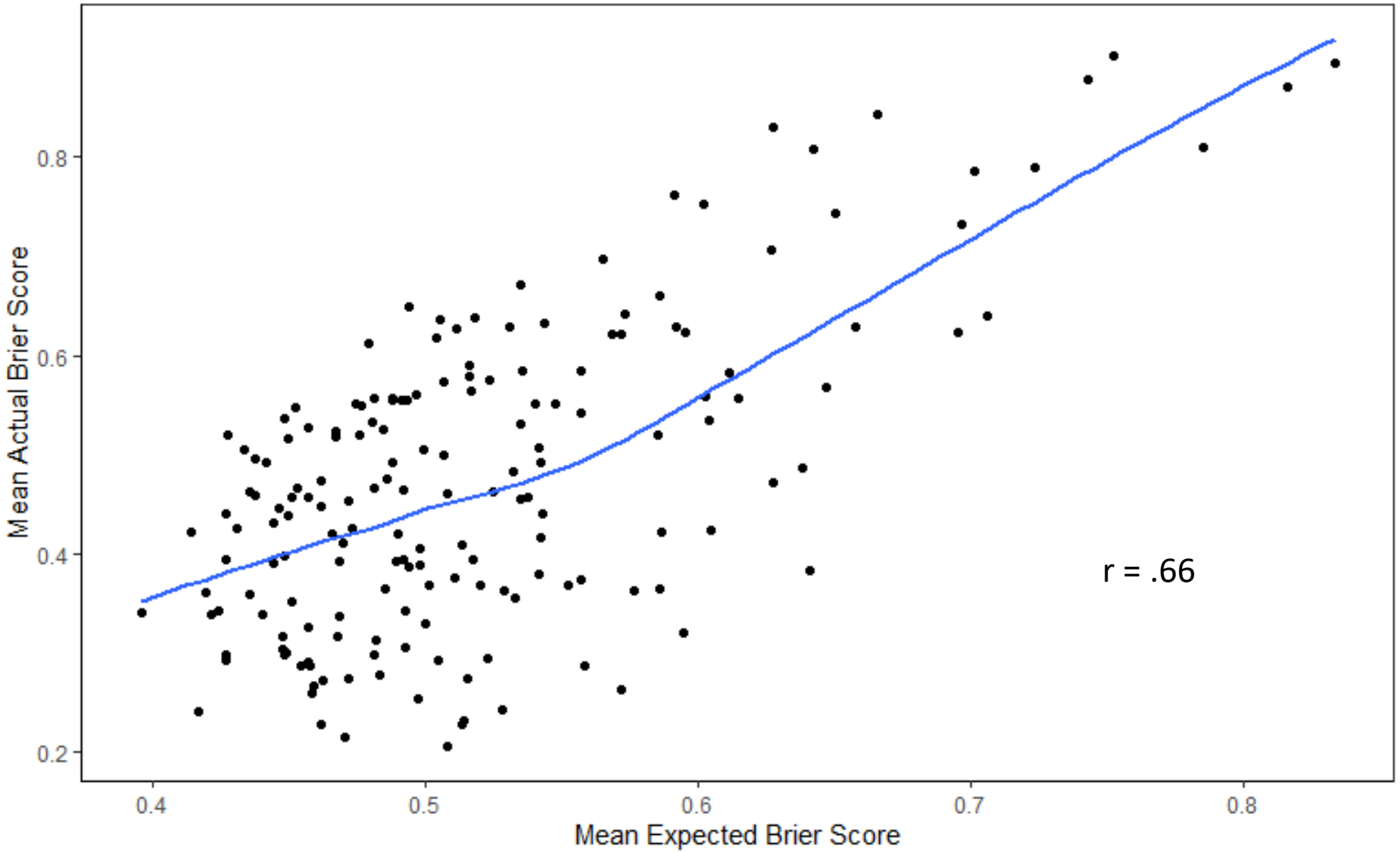


Overall r = .63

Slight elbows suggest poor performers are better discriminated than strong performers.

May also indicate reliably strong performers aren't sufficiently discriminated by sample size.

Overall Mean Correlations between Forecasters' Mean EBS and Mean BS



Elbow still shows up, but less pronounced

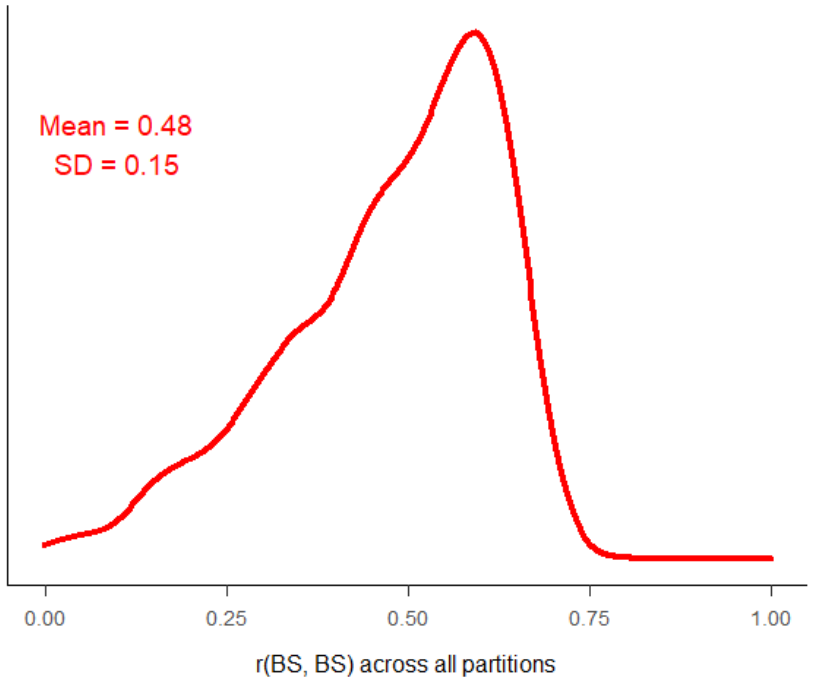
Research Question 1 answer: EBS and BS are highly correlated, particularly in the aggregate

Cross Validation

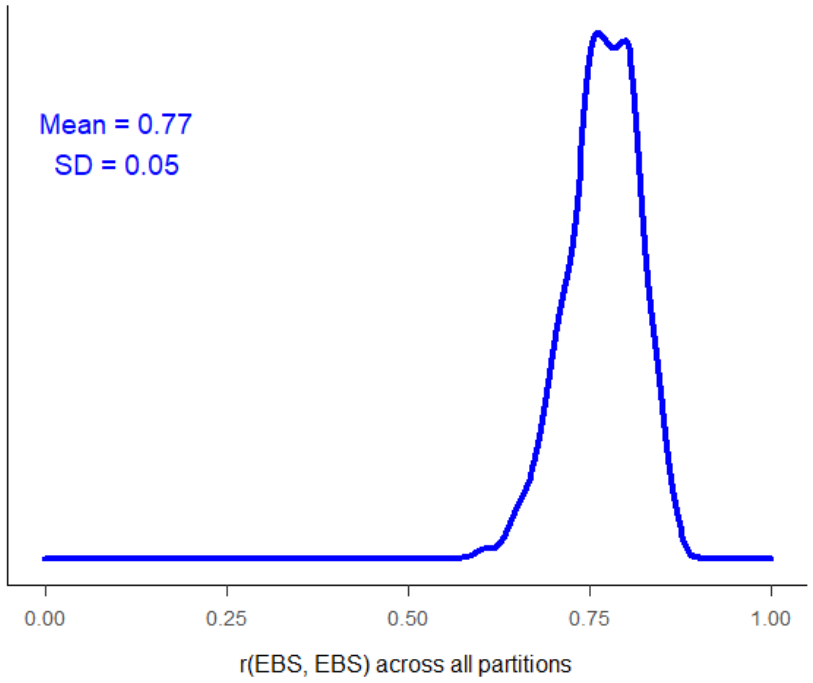
- There are 462 unique ways to group the 11 items into one group of 5 and another of 6
- For each partition, we calculated mean EBS and mean BS for each judge

Cross Validation: How well does EBS predict BS out-of-sample

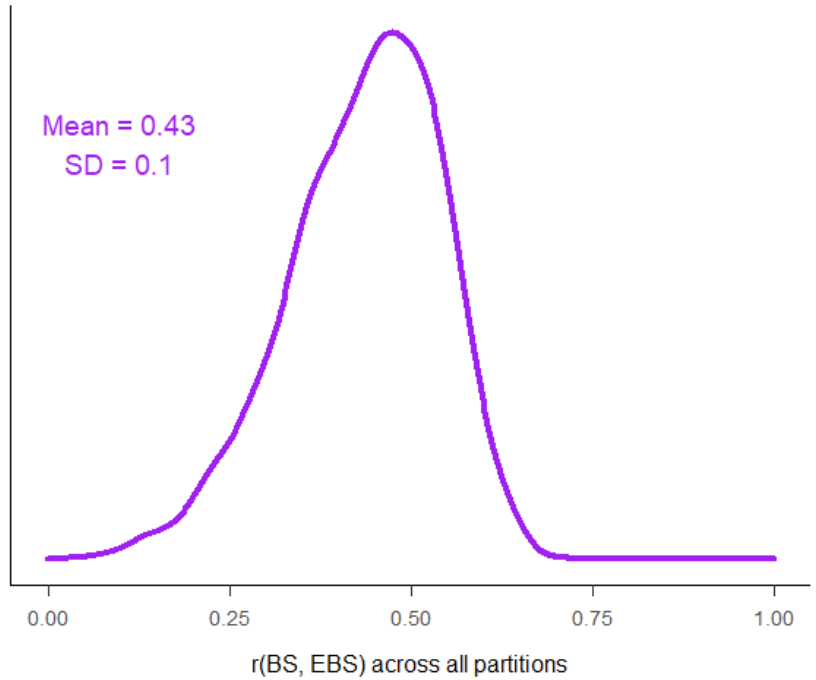
BS Reliability



EBS Reliability



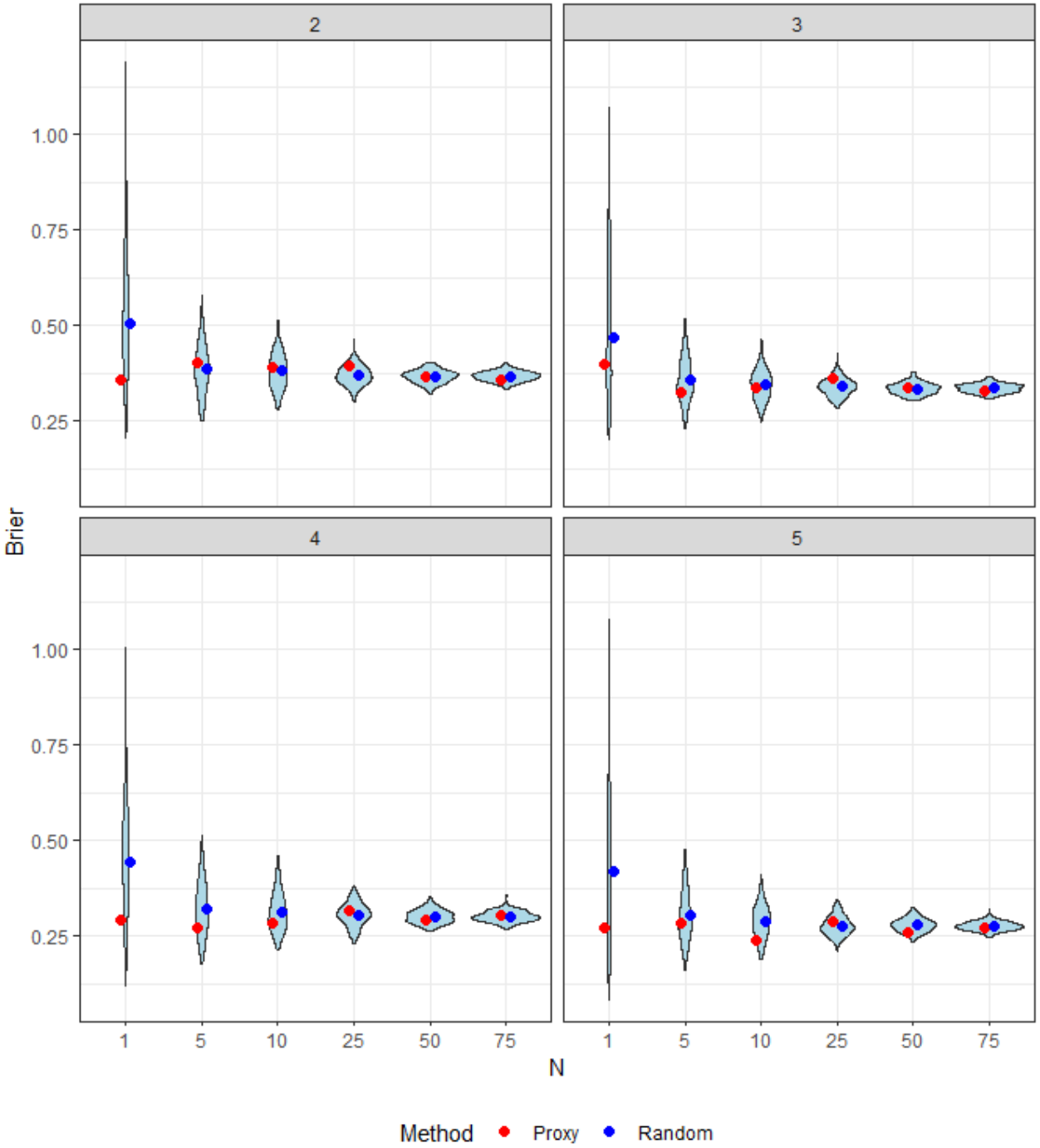
Correlation between EBS and BS



Research Question #2: EBS from one set of question predicts BS on another set almost as well as knowing BS itself from the first set would

Bootstrap Analysis: Wisdom of the Expectedly Accurate

- For each round rank all forecasters by **EBS performance on all rounds to date**
- In each subsequent round, compare **averaged crowd forecast** from **top ranked EBS performers** from previous rounds to **200 randomly selected samples** of the same size
  - For N = 1, distribution represents distribution of all forecasters (not randomly selected subsamples)
- Plotted distribution of mean aggregate Brier from **bootstrapped samples**, to compare with performance of **EBS selected sample**



Single forecasters selected by EBS rank:

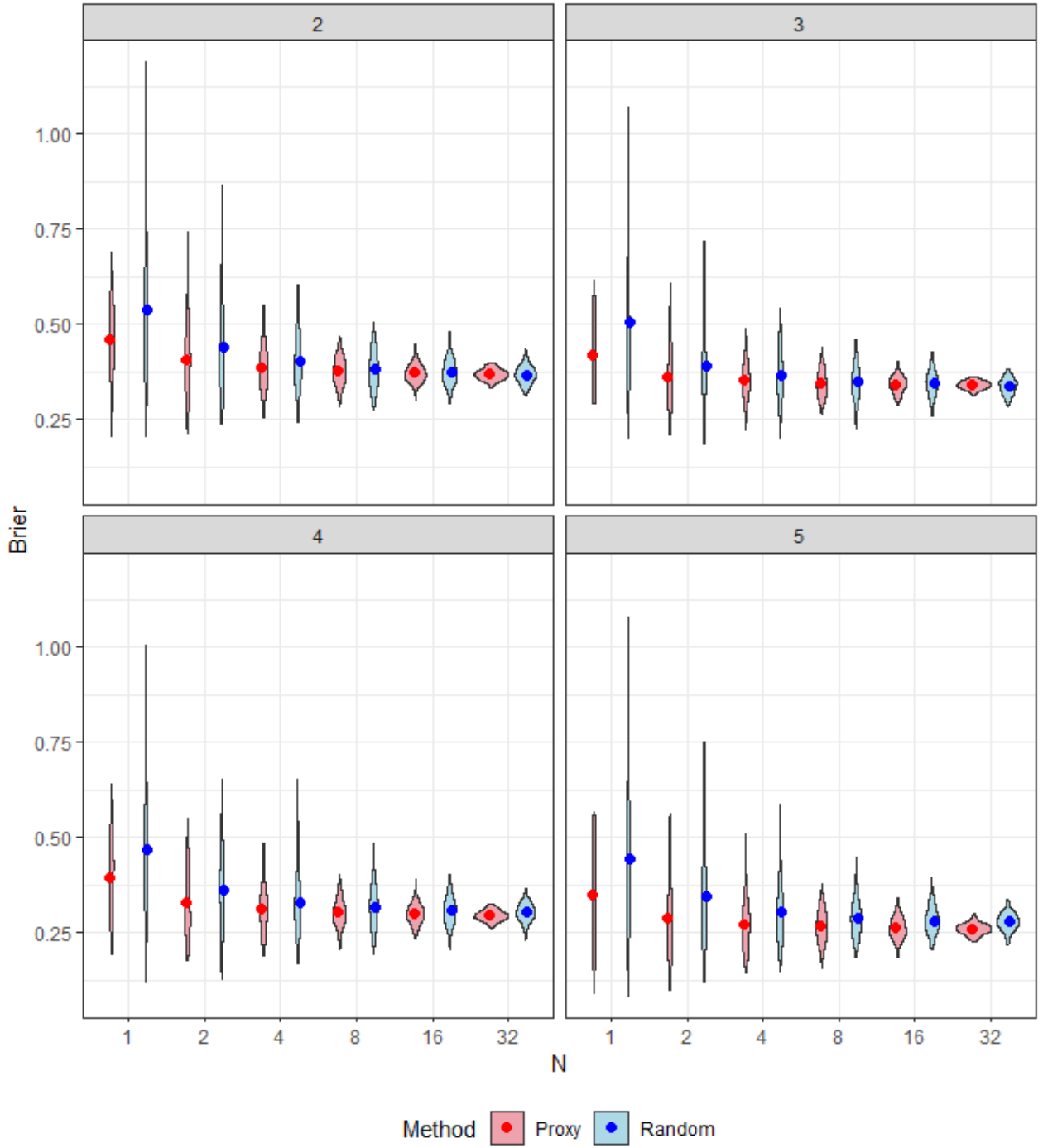
- Dramatically outperformed most individual forecasters
- Provided most of the benefits of the Wisdom of Crowds—were roughly as accurate as aggregations of much larger samples

No clear difference between high EBS performers and randomly selected larger samples

Strong vs Poor Expected Brier Scorers

#1 EBS performer was the same in every round. Perhaps they were selected by chance.

What if we randomly select forecasters either from full sample or from among the top 50 EBS performers?



EBS selection works by reducing chances individual forecasters make large errors

Larger random samples also eliminate large errors, but require more forecasters

Largest benefits of EBS selection were for smaller samples. **N = 1 improvement was not due to random chance**

Research Question #3: EBS allows us to identify accurate forecasters before any forecasting questions have resolved

Research Question #4: EBS does not make the crowd wiser, but it does allow us to extract the benefits of crowd wisdom in much smaller sample

Questions Answered

- How well do Proxy Scores correlate with actual accuracy?
  - **Very well, r = .66 between forecasters' average BS and EBS**
- Do Proxy Scores from one set of questions predict actual accuracy on other questions?
  - **Yes, EBS from one set of questions predicts BS on another almost as well as BS itself from the first set would**
- Can Proxy Scores identify high performing forecasters without knowing anything about their actual accuracy?
  - **Yes, EBS is both reliable and reliably predicts who the most accurate forecasters would be without knowing in advance, though its main benefit is filtering out error prone forecasters**
- Can Proxy Scores help us improve on current Wisdom of Crowds methods?
  - **They don't improve crowd accuracy, but allow us to achieve a similar level of accuracy with a much smaller sample, or even just a single high performing Proxy Scorer**

References

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*(7279), 294-295.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19-35.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267-280.

Himmelstein, M., Atanasov, P., & Budescu, D. V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment & Decision Making*, *16*(2).

Ho, E. (2020). *Developing and Validating a Method of Coherence-based Judgment Aggregation* [Unpublished doctoral dissertation]. Fordham University.

Liu, Y., Wang, J., & Chen, Y. (2020, July). Surrogate scoring rules. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 853-871).

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., ... & Tetlock, P. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, *21*(1), 1.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Witkowski, J., Atanasov, P., Ungar, L. H., & Krause, A. (2017, February). Proper proxy scoring rules. In *Thirty-First AAAI Conference on Artificial Intelligence*.