Submitted to *Management Science*
manuscript MS-0001-1922.65

Bias, Information, Noise: The BIN Model of Forecasting

Ville A. Satopää
INSEAD, Fontainebleau, France, ville.satopaa@insead.edu

Marat Salikhov
Yale School of Management, New Haven, CT, United States

Philip E. Tetlock, Barbara Mellers
The Wharton School of the University of Pennsylvania, Philadelphia, PA, United States

A four-year series of subjective-probability forecasting tournaments sponsored by the U.S. intelligence community revealed a host of replicable drivers of predictive accuracy, including experimental interventions such as training, teaming, and tracking-of-talent. Drawing on these data, we propose a Bayesian model (BIN: Bias, Information, Noise) for disentangling the underlying processes that enable certain forecasters and forecasting methods to out-perform: either by tamping down bias and noise in judgment or by ramping up the efficient extraction of valid information from the environment. The BIN model reveals the dominant driver of performance enhancement to be noise reduction, though some interventions also reduce bias and improve information extraction. Even "debiasing training" designed to attenuate bias improved accuracy largely by tamping down noise. Organizations may often discover that the most efficient way to boost forecasting accuracy is to target noise.

Key words: Bayesian Statistics; Judgmental Forecasting; Partial Information; Shapley Value; Wisdom of Crowds

1. Introduction

Forecasters must often work under less-than-optimal conditions: too little or too much data as well as data of uncertain or varying reliability. They must make best guesses about whether an investment will yield a target return, an unusual tumor warrants surgery or an adversary is violating an arms-control treaty (Armstrong 2001, Kahneman 2011, Tetlock and Gardner 2016).

Errors in extracting predictive signals are inevitable – and from a statistical perspective, these errors can be decomposed into bias and noise. Bias reflects predictable error. For instance, certain individuals in certain situations might display a systematic tendency to make more false-positive

judgments (more disappointing investments, unnecessary operations, or unfounded accusations) or more false-negative judgments (missing more opportunities to profit, save lives, or call out cheaters). The research literature on biases is voluminous (Gilovich et al. 2002) but their key feature is that they are systematic. If we knew enough about the forecasters' cognitive strategies and value priorities, it should be possible, in principle, to predict the direction and approximate magnitude of their deviations from accuracy.

By contrast, noise is unpredictable, inherently nonsystematic error. No matter how much we know about the forecaster, it is impossible to anticipate the direction or magnitude of these deviations from the true signal. Kahneman et al. (2016) argue that the research literature on noise is much less developed than that on bias because noise is much more difficult for human observers, wired up to detect patterns, to accept. It is easy to construct causal explanations for bias that invoke propensities in the forecasters – hubris, rigidity, prejudice, favoritism – but noise defies causal categorization.

Separating noise from bias in probabilistic forecasts of a singular event is difficult, arguably impossible. Putting aside judgments of zero and one, there is always wiggle to argue over whether something improbable happened. However, if we have access to forecasters' predictions about multiple events, it becomes possible to disentangle expected levels of noise and bias – and to treat their relative influence on forecasts as an open empirical question. To this end, we introduce the BIN model, a Bayesian approach to decomposing forecasting accuracy into three components: bias, partial information, and noise. This is a joint model of two groups of forecasters, which we denote as control and treatment, and allows us to calculate many useful summary statistics, such as the posterior probabilities of the treatment improving bias, information, and/or noise.

The remainder of the article applies the model to data from a multi-year, geopolitical forecasting tournament to explore the mechanisms via which three experimental interventions – training, teaming and tracking of talent – improved forecasts (Mellers et al. 2014). The BIN model reveals that noise reduction plays the dominant role in driving the effectiveness of each intervention, even that of debiasing training. Noise is a pervasive obstacle to judgmental accuracy – and noise-reduction may be a more cost-effective method of boosting accuracy than widely appreciated (Kahneman et al. 2016).

2. Modeling Bias, Information, and Noise in Forecasts

In messy real-world situations, forecasters are bound to make mistakes – and misinterpret signals from the environment about what is likely to happen next. We decompose forecasters' signal-extraction process into three components: bias (defined as systematic deviations between forecasters' interpretation of signals and the true informational value of those signals – deviations that

can take the form of either over- or under-estimation of probabilities), partial information (defined as the informational value of the subset of signals that forecasters use – relative to full information that would permit forecasters to achieve perfect accuracy) and noise (defined as the residual variability, independent of the outcome). To illustrate, we start with a simple example.

Example 1. Consider a multi-round game in which the researcher flips a fair coin twice and forecasters predict the probability that both tosses are heads. We denote heads with H and tails with T and list the four equally likely outcomes: TT, HT, TH, and HH.

Imagine first a forecaster with zero bias and zero noise in his judgments. If the forecaster has no other information, the forecaster should view the four outcomes as equally likely and predict 0.25, the base rate, for the outcome, HH. Suppose now that the forecaster has partial information. The researcher tells him the outcome of the first toss before asking for his prediction. Now the forecaster constructs a revised prediction for HH. If the first toss is a T, the two-head sequence would be impossible, and the forecaster would predict 0.0 for HH. If the first toss is H, the forecaster would predict 0.5 for HH (the probability of a second H). So, over multiple rounds, the forecaster would predict 0 or 0.5, with equal probability, depending on the outcome of the first toss. This variation is neither noise nor bias. It is attributable to partial information that results in predictions that vary around the base rate of 0.25.

This example illustrates two key points. First, given that the forecaster's prediction is equally likely to be 0 or 0.5, it has mean 0.25 and variance 0.0625. Furthermore, the covariance between the prediction and outcome is 0.0625. The variance-covariance equality is no coincidence. If forecasters are rational Bayesian agents seeking to minimize a proper scoring function, such as the Brier score, then all variance in their predictions is driven by partial information and is equivalent to the covariance with outcomes (Satopää et al. 2016).

Second, this example illustrates how forecasters can make different predictions even if there is zero bias or noise in their judgments. A forecaster with no information (no chance to see outcome of first toss) always predicts the base rate. A forecaster with partial information (knows outcome of first toss) predicts 0.0 or 0.5 for HH, depending on the outcome. A forecaster with complete information (who observes both tosses) could, of course, "predict" all outcomes.

Consider now more realistic cases in which bias and noise occur. Suppose a biased but noise-free forecaster thinks the fair coin is unfair, and the probability of heads is 0.6, thus over-predicting heads. If the forecaster has no partial information, he should always predict $0.6 \times 0.6 = 0.36$ for HH, over-shooting the base rate of 0.25. If this forecaster observes the first toss, he would predict 0.0 or 0.6 for HH, depending on whether the first toss is a T or H. The average of his predictions is now 0.3, which, again, exceeds the base rate.

Lastly, suppose there is noise but no bias in the forecaster's judgments. Specifically, the forecaster does not realize that the researcher has told him the result of an unrelated coin toss. Then, following our earlier argument, the forecaster predicts 0.0 or 0.5 for HH, depending on whether the unrelated toss comes out T or H, respectively. Given that this toss is independent of the two flips that determine the outcome, variability in the forecaster's predictions does not correlate with the outcome. Suppose now that the researcher tells the forecaster the outcome of two flips, only one of which actually determines the outcome. The forecaster then predicts 1.0 if the researcher reports HH; otherwise, he predicts 0.0. The mean and variance of the forecaster's prediction are 0.25 (from $0.25 \times 1 + 0.75 \times 0$) and 0.1875 (from $0.25 \times (0.25-1)^2 + 0.75 \times (0.25-0)^2$), respectively. Given that his average prediction equals the base rate, the forecaster is unbiased. Not all variability in his predictions, however, covaries with outcomes. There are 8 equally-likely cases. In 5/8 of the cases the forecaster correctly predicts 0.0; in 1/8 cases he correctly predicts 1.0; and in the remaining 2/8 cases his prediction does not match the outcome. Therefore the covariance between his prediction and the outcome is $(5/8) \times (0.25-0) \times (0.25-0) + (1/8) \times (0.25-1) \times (0.25-1) + (2/8) \times (0.25-1) \times (0.25-0) = 0.0625$, which aligns with our earlier noise-free and unbiased forecaster who also observed the first toss. Therefore, in this example, partial information is estimated to be 0.0625 and noise is 0.125 (from 0.1875–0.0625).

Although each of the three components – bias, noise, and (inevitably partial) information – has an intuitive definition, one can imagine a variety of models that capture their effects. At a minimum, though, one's model must describe two comparison groups of forecasters jointly. Only then can we can make "significance" statements about common mechanisms via which the treatment affects accuracy. For instance, we can estimate the probability that the treatment group is 20% more accurate than the control group and that 30% of that differential is attributable to less bias. Such comparisons would not be possible if the two groups were analyzed independently with separate models. In short, we need a joint probability model to make joint statements about groups.

Unfortunately, no standard distribution can jointly describe a binary outcome, which can only take on values of zero or one, and multiple probabilities, which are bounded by zero and one. The solution we adopt here is to describe the forecasting process with flexible well-known distributions, such as the multivariate normal.

To make this specific, we follow Satopää et al. (2016) and posit a Signal Universe that contains all signals of positive or negative relevance to event occurrence. In this realm, the event happens if and only if the combined contribution of all relevant signals is positive. Borrowing then from the statistical theory of generalized linear models (McCullagh 2019) and probit regression (Bliss 1934) in particular, we model the binary event with a hypothetical continuous variable, the accumulation of all relevant signals along which the event happens if and only if the variable is positive.

In addition to relevant signals, the universe contains signals that have zero relevance and are uncorrelated with the outcome. Forecasters sample and interpret signals with varying skill and thoroughness. They may sample relevant signals (increasing partial information) or irrelevant signals (creating noise). Furthermore, they may center the signals incorrectly (creating bias). The accumulation of these signals is then modeled with continuous variables that exhibit varying degrees of bias, partial information, and noise in their forecasts. These variables summarize forecasters' (often noisy and biased) interpretations about how the signals they have observed relate to the outcome.

Modeling outcomes and forecasters with continuous variables provides three benefits: (a) we can capture partial information as the covariance between forecasters' interpretations and the outcome-determining variable; (b) when we introduce mean-zero noise into forecasters' interpretations, the variance of forecasts increases but not the mean or partial information of signal interpretation; (c) when we introduce bias into forecasters' interpretations, the mean of forecasts changes, but not the variance or partial information of signal interpretation. Bias and noise thus play disjoint roles in forecasters' judgments.

In statistical terms, the outcome-determining variable and the forecasters' interpretations of signals are latent variables because they are not directly observed by the experimenter. Our first modeling assumption concerns the distribution of these variables.

ASSUMPTION 1. *The latent variables that determine the outcome and the forecasters' interpretations of signals are normally distributed.*

Even though real-world data rarely follow the normal distribution exactly, this is often a reasonable approximation and commonly assumed in statistical methodology. For instance, probit regression, which served as our starting point, assumes normally distributed latent variables. In our context, if each forecaster interprets a large number of independent signals and the signals have small tails, then the central limit theorem justifies our assumption of normality. As a check, we performed sensitivity analyses by applying our BIN model to latent variables that were simulated from the (multivariate) t-distribution. The estimation process could recover the true parameters with reasonable accuracy as long as the tails of the latent variables were not very heavy, suggesting low sensitivity to the exact distribution of the latent variables. For more details, see the supplementary material.

Delving into the technical details of the BIN model, we consider first a single forecaster predicting an unknown event. Denote this event with $Y \in \{0, 1\}$ such that $Y = 1$ if the event happens and $Y = 0$ if the event does not. The outcome is determined by a hypothetical normally distributed variable Z^* such that $Y = \mathbf{1}(Z^* > 0)$, where the indicator function $\mathbf{1}(E)$ equals 1 if E is true;

otherwise, 0. The expected frequency of this event must align with a given base rate, $p^* \in (0,1)$, which we can do, without loss of generality, by fixing $\text{Var}(Z^*) = 1$ and choosing the mean $\mu^* = \mathbb{E}[Z^*]$ such that $\mathbb{P}(Z^* > 0) = p^*$, the base rate. We then have:

$$p^* = \mathbb{P}(Y = 1) = \mathbb{P}(Z^* > 0) = 1 - \mathbb{P}(Z^* \leq 0) = 1 - \Phi(-\mu^*) = \Phi(\mu^*),$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal distribution. Inverting this function gives us: $\mu^* = \Phi^{-1}(p^*)$.

A forecaster has a probability $p_0 \in (0,1)$ for the event $\{Z^* > 0\}$ based on a normally distributed variable Z_0 that represents his interpretation of the signals. The variable Z_0 describes the forecaster's bias, noise, and partial information.

The more Z_0 covaries with Z^*, the more information the forecaster has about the event. If Z_0 and Z^* are perfectly correlated, a forecaster with no noise or bias can deduce the value of Z^* and perfectly predict the event. More frequently, the forecaster must work with partial information. For instance, forecasters may closely follow news about British politics, which will strengthen their signals. But it will not let them predict Brexit with certainty. Irreducible uncertainty will remain. Following Satopää et al. (2016), we introduce partial information in the BIN model with the parameter $\text{Cov}(Z_0, Z^*) = \gamma_0$. The greater γ_0, the more Z_0 covaries with Z^* and the more accurate the forecaster.

Given that both Z_0 and Z^* are on continuous scales, bias equals the difference between the means whereas noise equals any variability in Z_0 that does not covary with Z^*. If the mean of the forecaster's interpretation is $\mathbb{E}[Z_0] = \mu^* + \mu_0$, then bias is $\mathbb{E}[Z_0] - \mathbb{E}[Z^*] = \mu_0$. Noise is the remaining variability of Z_0 after removing all covariance with Z^*: $\text{Var}(Z_0) - \text{Cov}(Z^*, Z_0) = \delta_0$.

To summarize, Z_0 and Z^* follow a multivariate normal distribution:

$$\begin{pmatrix} Z^* \\ Z_0 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu^* \\ \mu^* + \mu_0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma_0 \\ \gamma_0 & \gamma_0 + \delta_0 \end{pmatrix} \right),$$

where

$$\begin{aligned} \text{Outcome:} \quad & Y = \mathbf{1}(Z^* > 0) \\ \text{Bias:} \quad & \mu_0 = \mathbb{E}[Z_0] - \mathbb{E}[Z^*] \\ \text{Information:} \quad & \gamma_0 = \text{Cov}(Z_0, Z^*) \\ \text{Noise:} \quad & \delta_0 = \text{Var}(Z_0) - \text{Cov}(Z_0, Z^*) \end{aligned}$$

The bias μ_0 can take on any value between negative and positive infinity, and causes the forecaster's interpretation Z_0 to be either too high ($\mu_0 > 0$) or too low ($\mu_0 < 0$); noise δ_0 ranges from no noise ($\delta_0 = 0$) to infinite noise ($\delta_0 = \infty$); and partial information varies from no information ($\gamma_0 = 0$)

to perfect information ($\gamma_0 = 1$). The partial information parameter is bounded by 1 because it represents co-variability with Z^* that only has variance 1. If the expert is unbiased ($\mu_0 = 0$), noise-free ($\delta_0 = 0$), and has perfect information, then $Z_0 = Z^*$ and $\gamma_0 = \text{Var}(Z_0) = \text{Var}(Z^*) = 1$.

Given that the forecaster reports the probability of the event $\{Y = 1\}$, not the interpretation Z_0, the model should describe how the forecaster converts the interpretation into a probability prediction. The rational Bayesian belief of Y given Z_0 is $\mathbb{E}[Y|Z_0] = \mathbb{P}[Z^* > 0|Z_0]$. The standard results on the conditional distributions of normal random variables (e.g., Ravishanker and Dey 2001) show:

$$Z^*|Z_0 \sim \mathcal{N}\left(\mu^* + \frac{\gamma_0}{\gamma_0 + \delta_0}\left(Z_0 - \mu^* - \mu_0\right), 1 - \frac{\gamma_0^2}{\gamma_0 + \delta_0}\right).$$

Thus,

$$\mathbb{P}[Z^* > 0|Z_0] = 1 - \Phi\left(-\frac{\mu^* + \frac{\gamma_0}{\gamma_0+\delta_0}\left(Z_0 - \mu^* - \mu_0\right)}{\sqrt{1 - \frac{\gamma_0^2}{\gamma_0+\delta_0}}}\right) = \Phi\left(\frac{\mu^* + \frac{\gamma_0}{\gamma_0+\delta_0}\left(Z_0 - \mu^* - \mu_0\right)}{\sqrt{1 - \frac{\gamma_0^2}{\gamma_0+\delta_0}}}\right), \quad (1)$$

where the last step follows from the symmetry of the normal distribution. Specifically, $\Phi(a)$ is the area under the probability density function of a standard normal and to the left of a. Given that the standard normal distribution is symmetric about zero, this area is equal to the area to the right of $-a$, namely $1 - \Phi(-a)$. Therefore $\Phi(a) = 1 - \Phi(-a)$, which is the equality used in the last step of equation (1).

Unfortunately, forecasters' bias and noise are not identifiable from the probability forecasts in equation (1). This perfect-rationality equation assumes that forecasters know the level of bias and noise in their forecasts and automatically self-correct so their final conditional probabilities $\mathbb{P}[Z^* > 0|Z_0]$ exhibit zero noise or bias. Indeed, the forecasts in (1) are equal (in distribution) to unbiased and noise-free forecasts with $\frac{\gamma_0^2}{\gamma_0+\delta_0}$ amount of information. Given that we cannot statistically distinguish between this case and the original setting in (1), where the forecasts were corrected for bias and noise, the parameter values cannot be identified.

To identify the components, however, we only need to make the plausible bounded-rationality assumption that the forecaster is not aware of the noise and bias in the interpretation Z_0 – and believes that $\delta_0 = 0$ and $\mu_0 = 0$. Plugging in this potential for misbeliefs into (1), the forecaster now predicts:

$$p_0 = \Phi\left(\frac{Z_0}{\sqrt{1 - \gamma_0}}\right).$$

The resulting probability prediction can exhibit both bias and noise – and allows us to use the predictions to study the bias, noise, and partial information in the forecaster's judgments.

Overall, this construction aligns with the more familiar situation in which forecasters make point predictions about a continuous outcome, such as the exact change in inflation or national GDP. Bias here would be any difference in the average outcome and prediction, and noise is any uncorrelated variability in the predictions. In such situations, it is reasonable to posit that forecasters are not aware of their bias or noise: if they were, they could improve their accuracy by reducing them.

We can extend the BIN model to groups of multiple forecasters, which we call "control" and "treatment" conditions and designate their predictions and components of accuracy with subscripts of 0 and 1, respectively. As before, each forecaster bases a prediction on different signals about Z^*, and the expected levels of bias, noise, and information are described by the model parameters. Allowing each forecaster to have a different set of parameters would lead to parameter proliferation. Ideally we want only one bias, noise, and information parameter per group of forecasters, which we can achieve by treating all forecasters of the same type or group symmetrically and as exchangeable.

Assumption 2. *Forecasters within each group are exchangeable.*

Under exchangeability, all forecasters in the same group can make different forecasts based on different interpretations but have the same expected level of bias, noise, and information. Denote the expected levels of bias, information, and noise for forecasters in the control and treatment groups with $(\mu_0, \gamma_0, \delta_0)$ and $(\mu_1, \gamma_1, \delta_1)$, respectively. These model parameters capture average behavior within each group. Although this prevents statements at the individual-forecaster level, it allows us to consider different sets of forecasters per event and different numbers of forecasters predicting each event. This is useful because, due to fatigue, time constraints, and self-selection, forecasters rarely predict exactly the same events at the same time. By assuming exchangeability we gain data and statistical power.

Specifically, the jth forecaster in group $g \in \{0,1\}$ predicts $Y = \mathbf{1}(Z^* > 0)$ based on the interpretation $Z_{g,j} \sim \mathcal{N}(\mu^* + \mu_g, \gamma_g + \delta_g)$, and $\text{Cov}(Z_{g,j}, Z^*) = \gamma_g$. A summary of our final BIN model then is:

$$
\begin{aligned}
&\text{Outcome:} && Y = \mathbf{1}(Z^* > 0) \\
&\text{Control Group:} && \begin{cases} \text{Bias:} & \mu_0 = \mathbb{E}[Z_{0,j}] - \mathbb{E}[Z^*] \\ \text{Information:} & \gamma_0 = \text{Cov}(Z_{0,j}, Z^*) \\ \text{Noise:} & \delta_0 = \text{Var}(Z_{0,j}) - \text{Cov}(Z_{0,j}, Z^*) \\ j\text{th forecaster's prediction:} & p_{0,j} = \Phi\left(\frac{Z_{0,j}}{\sqrt{1-\gamma_0}}\right) \end{cases} \\
&\text{Treatment Group:} && \begin{cases} \text{Bias:} & \mu_1 = \mathbb{E}[Z_{1,j}] - \mathbb{E}[Z^*] \\ \text{Information:} & \gamma_1 = \text{Cov}(Z_{1,j}, Z^*) \\ \text{Noise:} & \delta_1 = \text{Var}(Z_{1,j}) - \text{Cov}(Z_{1,j}, Z^*) \\ j\text{th forecaster's prediction:} & p_{1,j} = \Phi\left(\frac{Z_{1,j}}{\sqrt{1-\gamma_1}}\right) \end{cases}
\end{aligned}
$$

So far we have described a model for a single event. We extend the BIN model to multiple events by assuming that predictions and outcomes are independent and identically distributed across events. This standard assumption is often a reasonable approximation of reality and promotes parsimony in modeling. For instance, the standard probit regression, which served as our starting point, relies on this assumption.

ASSUMPTION 3. *Outcomes and predictions are independent and identically distributed across events.*

This assumption has two parts: a) the base rate and the forecasters' expected levels of bias, noise, and partial information are the same across events; and b) conditional on these values, predictions and outcomes are independent across events. Put differently, the outcome and forecasters' predictions for one event provide no additional information about the outcome and predictions for any other event – as long as we know the base rate and the parameters representing the forecasters' bias, noise, and information.

Consider, for instance, forecasters with low bias, low noise, and high information – a profile conducive to accuracy. Conditional on this profile, knowing these forecasters' predictions were 25% for, say, German-Spanish bond yield spreads tells us nothing about their predictions about, say, the Syrian civil war. All we can say is that they are likely to make predictions that are directionally accurate and relatively extreme (closer to 0.0 and 1.0). The basis for this claim, however, is knowledge of their high information and low noise and bias, not of their predictions for an earlier event.

As a sensitivity analysis, we tested our model on simulated data with varying degrees of dependence across outcomes. As the correlation approaches 1.0, outcomes provide no more information about the parameters than would a single outcome. Therefore, as outcome inter-dependency grows, we need more data to maintain a constant level of estimation accuracy. Fortunately, with the amount of data from our tournaments, we recovered the true parameter values with reasonable accuracy as long as the correlation among the outcomes did not exceed 0.5. These results can be found in the supplementary material.

A correlation above 0.5 corresponds to a high inter-dependency among the outcomes and is unlikely to occur in our data because IARPA (the sponsor of the forecasting tournament) chose the forecasting questions and sought to measure the forecasting equivalent of fluid intelligence, as uncontaminated as possible by crystallized intelligence—or specialized knowledge of persons, places or political processes. To this end, they chose extraordinarily diverse questions: from German-Spanish bond yield spreads to the Syrian civil war to island building in the South China Sea to

Arctic sea-ice mass to Ebola epidemics. Given that these outcomes are unlikely to cluster or show much dependence, the independence part of Assumption 3 seems reasonable.

To measure dependence across events, we calculated the correlation between the predictions of one event and the predictions of another event within each experimental condition in our data. We excluded a pair of events whenever the number of forecasters fell below 50, which left us with more than $1,200$ pairs of events per condition. Across conditions, the average and median correlation were 0.12 and 0.13, respectively. The inter-quartile range of the correlations was $[0.0, 0.25]$. This suggests that, on average, Assumption 3 is a fair approximation for our data. Although there were mild positive or negative correlations, 97% of the correlations were within the safe range of our sensitivity analysis $[-0.5, 0.5]$.

Although in our data, outcomes and predictions are likely to be uncorrelated across events, predictions of forecasters are likely to be correlated within each event. The predictions target the same outcome, and hence their correlation can stem from shared information as well as shared misconceptions among forecasters. Our model does not make assumptions about the magnitude of the within-event correlation but instead treats it as an empirical matter. Therefore the predictions of the same outcome can be dependent or independent, based on what is supported by the data. Furthermore, we model the within-event dependence separately for each group and across the two groups. For example, the dependence within the control group does not need to equal the dependence in the treatment group or the dependence between forecasters from different groups. Given that these parameters are not of direct interest here, we defer their discussion to the supplementary material.

Taken together, Assumption 2 and 3 mean that our parameters should capture the average levels of bias, noise, and information within each group of forecasters and across questions. The assumptions are reasonable in our case because the goal is to understand groups' average behavior and make general, not question-specific, statements about bias, information, and noise.

3. Data, Methods, and Results

Tables 1-3 apply the BIN model to the dataset for the geopolitical forecasting tournament sponsored by IARPA in 2011-2015. This section describes the data and methods used in our analysis.

3.1. Data

The dataset includes hundreds of forecasting questions, outcomes, and probabilistic predictions by the thousands of participants in the Good Judgment Project (GJP). For instance, one question asked whether Serbia would be granted EU candidacy by 31 December 2011. Forecasting began on September 1, 2011. The question resolved as "no" because Serbia did not gain EU candidacy by the target date. The question was thus open for 4 months. All GJP data are publicly available.[1]

[1] The data can be downloaded at `https://goodjudgment.io/gjp/gjp_data.zip`.

Table 1 Parameter estimates (with 95% CI) and posterior inference for the Good Judgment Project data. The first two columns show effects of probability training among individuals and in teams. The next two columns represent effects of teaming in untrained and trained individuals. The last two columns show the effects of tracking among non-trained individuals and those who are already trained and working in teams.

	Training		**Teaming**		**Tracking**	
	Individuals: untrained vs. trained	**Teams: untrained vs. trained**	**Untrained: indiv. vs. teams**	**Trained: indiv. vs. teams**	**Trained: teams vs. supers**	**Untrained indiv. vs. supers**
Parameter estimates (with 95% CI)						
Outcome mean: μ^*	-0.90 [-1.10, -0.68]	-0.92 [-1.23, -0.60]	-0.94 [-1.24, -0.65]	-0.88 [-1.09, -0.69]	-0.84 [-1.09, -0.60]	-0.84 [-1.10, -0.60]
Bias (control): μ_0	0.46 [0.28, 0.65]	0.41 [0.12, 0.74]	0.52 [0.23, 0.80]	0.43 [0.26, 0.61]	0.22 [-0.01, 0.45]	0.36 [0.14, 0.60]
Bias (treatment): μ_1	0.45 [0.26, 0.64]	0.40 [0.12, 0.70]	0.44 [0.16, 0.73]	0.31 [0.16, 0.48]	-0.09 [-0.30, 0.13]	-0.09 [-0.29, 0.12]
Diff. in bias: $\lvert\mu_0\rvert - \lvert\mu_1\rvert$	0.01 [-0.01, 0.04]	0.02 [-0.03, 0.07]	0.07 [0.01, 0.13]	0.11 [0.07, 0.15]	0.10 [-0.27, 0.35]	0.25 [-0.13, 0.53]
Info. (control): γ_0	0.34 [0.27, 0.40]	0.29 [0.12, 0.41]	0.21 [0.06, 0.33]	0.31 [0.23, 0.38]	0.44 [0.34, 0.52]	0.33 [0.23, 0.41]
Info. (treatment): γ_1	0.34 [0.28, 0.40]	0.34 [0.20, 0.44]	0.31 [0.17, 0.42]	0.45 [0.39, 0.51]	0.60 [0.52, 0.66]	0.60 [0.52, 0.66]
Diff. in info.: $\gamma_0 - \gamma_1$	-0.00 [-0.03, 0.03]	-0.05 [-0.12, 0.00]	-0.10 [-0.17, -0.03]	-0.14 [-0.18, -0.10]	-0.16 [-0.22, -0.10]	-0.27 [-0.34, -0.20]
Noise (control): δ_0	0.91 [0.76, 1.08]	0.82 [0.53, 1.25]	1.09 [0.79, 1.49]	0.77 [0.62, 0.96]	0.64 [0.46, 0.93]	1.01 [0.79, 1.30]
Noise (treatment): δ_1	0.70 [0.57, 0.85]	0.53 [0.34, 0.85]	0.76 [0.51, 1.11]	0.54 [0.41, 0.69]	0.27 [0.15, 0.45]	0.28 [0.15, 0.48]
Posterior inferences						
Less bias in treatment group: $\mathbb{P}(\lvert\mu_1\rvert < \lvert\mu_0\rvert)$	0.86	0.77	0.99	1.00	0.72	0.91
Less noise in treatment group: $\mathbb{P}(\delta_1 < \delta_0)$	1.00	1.00	1.00	1.00	1.00	1.00
More information in treatment group: $\mathbb{P}(\gamma_0 < \gamma_1)$	0.53	0.97	1.00	1.00	1.00	1.00

To be included in our analysis, a question had to satisfy two criteria: (i) a binary outcome (yes/no); (ii) be open no more than 180 days. This makes the problems more comparable. To ensure forecaster bias has a consistent interpretation, we standardized the orientation of the outcomes by rescoring the original questions so that "yes" always refers to change from the status quo. Forecasters are thus predicting probabilities of change, and bias is either systematic over- or under-estimation of change.

Forecasters were encouraged to update predictions when their beliefs changed. If forecasters did not update on a given day, it was assumed their beliefs had not changed. On any given day, we treated forecasters' most recent forecasts as their up-to-date beliefs about the event. Given that our model produces many results on any given day, it is impractical to analyze all possible time horizons in depth. Therefore, as a special case, we present detailed results on predictions that were made 30 days prior to outcome resolution. This puts all forecasters on the same temporal-distance "playing field." And it ensures at least some uncertainty at the time of their predictions. But we also repeat the analysis for each of the 60 days before resolution dates and summarize those

Table 2 Analysis of predictive performance for the Good Judgment Project data.

	Training		Teaming		Tracking	
	Individuals: untrained vs. trained	**Teams: untrained vs. trained**	**Untrained: indiv. vs. teams**	**Trained: indiv. vs. teams**	**Trained: teams vs. supers**	**Untrained indiv. vs. supers**
Predictive performance						
Actual Brier score (control)	0.21	0.18	0.22	0.19	0.14	0.19
Actual Brier score (treatment)	0.19	0.16	0.18	0.14	0.08	0.08
Contributions						
Value of the contribution						
Reduction in bias	0.00	0.00	0.01	0.01	0.02	0.04
Increase in information	0.00	0.00	0.01	0.01	0.01	0.02
Reduction in noise	0.01	0.02	0.02	0.02	0.03	0.06
Percentage of control group Brier score						
Reduction in bias	0.8%	0.9%	3.8%	6.0%	9.9%	15.1%
Increase in information	0.0%	1.3%	2.0%	3.8%	6.5%	8.0%
Reduction in noise	6.2%	9.9%	8.2%	8.2%	17.2%	23.4%
Maximum achievable contribution						
Reduction in bias	33.1%	30.0%	34.5%	31.2%	16.4%	24.6%
Increase in information	15.9%	19.3%	15.1%	19.5%	26.9%	20.1%
Reduction in noise	50.3%	50.1%	49.8%	48.7%	55.9%	54.6%

Table 3 Summary statistics for the Good Judgment Project data.

	Training		Teaming		Tracking	
	Individuals: untrained vs. trained	**Teams: untrained vs. trained**	**Untrained: indiv. vs. teams**	**Trained: indiv. vs. teams**	**Trained: teams vs. supers**	**Untrained indiv. vs. supers**
Data summary						
Number of questions	191	87	87	191	140	140
Median size of control group	75	68	116	54	70	67
Median size of treatment group	54	77	68	68	52	52

findings. To avoid infinite probit scores, all predictions of exactly 0 or 1 were transformed to 0.0001 and 0.9999, respectively.

To explore determinants of accuracy, the researchers randomly assigned forecasters to treatment groups:

- Probability training: Forecasters completed a tutorial on probabilistic reasoning, drawing on recommendations from the forecast-elicitation literature (O'Hagan et al. 2006). They learned to consider reference classes; average multiple predictions from different sources; and avoid judgmental biases (e.g., over-confidence, confirmation bias, base-rate neglect). Manipulation checks verified forecasters had mastered the content.
- Teaming: Forecasters worked to teams in which they could debate each other's predictions.
- Tracking: Forecasters' performances were tracked over time. At the end of each tournament year, the top 2% forecasters were designated "superforecasters" and given an opportunity to work together next year (Mellers et al. 2015). We call this intervention tracking because of its

resemblance to educational policies in which children with similar abilities are placed in the same classroom.

Our goal is to study how these treatment conditions influence bias, noise, and information by comparing treatment groups to control conditions.

3.2. Methods

We estimate model parameters with Bayesian statistics that treat parameters (such as μ^*, μ_0, and so on) as random variables. With any Bayesian model, we need two components: the *prior* distribution of the parameters, that captures the researcher's uncertainty about parameters before observing the data, and the *likelihood*, that specifies the probability of the data as a function of the parameters. We then use Bayes' rule to update the prior distribution in light of the observed data. The updated distribution, known as the *posterior*, describes all uncertainty in the parameters after accounting for researchers' prior beliefs and data.

Section 2 describes the likelihood function (with technical details in Appendix A). For the prior, we use a uniform distribution that posits all parameter configurations across the two types of forecasters to be a priori equally likely. With this assumption, we are saying the prior distribution has minimal impact on the final posterior inference and the data drive the final results. Typically the posterior distribution cannot be derived analytically. Instead, the posterior distribution is estimated with Markov Chain Monte Carlo (MCMC) methods. The final output is a sample from the joint posterior distribution of the parameters.

We use this output to compare the parameters for different types of forecasters. First, we provide the posterior means of the bias, noise, and information parameters. Second, to understand the uncertainty in the parameters, we give 95% credible intervals. Third, by comparing components within each draw of the posterior sample, we can give posterior probabilities to the parameter estimates indicating the likelihood of the treatment group out-performing the control group. Fourth, we calculate how much the treatments improve accuracy via changes in the expected bias, noise, and information.

The last point requires elaboration. In our model, the expected Brier score of a forecaster in group $g \in \{0,1\}$ is

$$\mathrm{BrS}(\mu_g, \delta_g, \gamma_g, \mu^*) = \mathbb{E}_{Y,Z_g}\left\{\left[Y - \Phi\left(\frac{Z_g}{\sqrt{1-\gamma_g}}\right)\right]^2\right\}, \tag{2}$$

where $Y = \mathbf{1}(Z^* > 0)$, $Z^* \sim \mathcal{N}(\mu^*, 1)$, $Z_g \sim \mathcal{N}(\mu^* + \mu_g, \gamma_g + \delta_g)$, and $\mathrm{Cov}(Z_g, Z^*) = \gamma_g$. We present the analytical expression of (2) and its derivation in Appendix B. To illustrate how the expected Brier score behaves as a function of bias, information, and noise, Figure 1 fixes $\mu^* = 0$ and presents

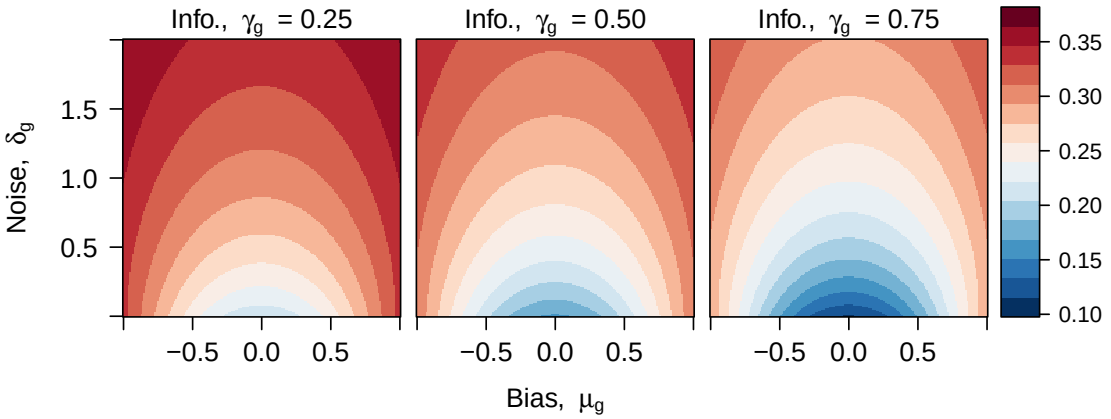

Figure 1 **Panels illustrating the expected Brier score under different values of bias, information, and noise. The base rate is fixed at 0.5.**

the expected Brier score for different combinations of μ_g, δ_g, and γ_g. The level of information γ_g changes from 0.25 in the left-panel to 0.50 in the middle and finally to 0.75 in the right panel. The x- and y-axis in each panel vary the levels of bias, μ_g, and noise, δ_g, over the ranges $[-1, 1]$ and $[0, 2]$, respectively. The expected Brier score is well behaved: it increases with bias and noise but decreases with information.

The expected treatment effect then is

$$\Delta\text{BrS} = \text{BrS}(\mu_0, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_1, \delta_1, \gamma_1, \mu^*).$$

Differences arise when $(\mu_0, \delta_0, \gamma_0)$ and $(\mu_1, \delta_1, \gamma_1)$ differ. Intuitively, the individual contribution of each parameter could be isolated by changing that parameter and observing the change in the expected Brier score. Specifically, the contributions would be

$$\begin{aligned}
\text{Bias:} &\qquad \text{BrS}(\mu_0, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_1, \delta_0, \gamma_0, \mu^*) \\
\text{Information:} &\qquad \text{BrS}(\mu_0, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_0, \delta_0, \gamma_1, \mu^*) \\
\text{Noise:} &\qquad \text{BrS}(\mu_0, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_0, \delta_1, \gamma_0, \mu^*)
\end{aligned}$$

Unfortunately, the parameters interact. The effect of any given parameter depends on the values of the other two parameters. This means the sum of the above contributions does not necessarily equal the sum of the overall change, ΔBrS.

To solve this problem, parameter changes must be computed sequentially. For instance, the contributions due to changing bias first, then information, and finally noise are

$$\begin{aligned}
\text{Bias:} &\qquad \text{BrS}(\mu_0, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_1, \delta_0, \gamma_0, \mu^*) \\
\text{Information:} &\qquad \text{BrS}(\mu_1, \delta_0, \gamma_0, \mu^*) - \text{BrS}(\mu_1, \delta_0, \gamma_1, \mu^*) \\
\text{Noise:} &\qquad \text{BrS}(\mu_1, \delta_0, \gamma_1, \mu^*) - \text{BrS}(\mu_1, \delta_1, \gamma_1, \mu^*)
\end{aligned}$$

These differences form a telescoping sum that equals the overall change ΔBrS. But the order of parameters matters. For instance, in the above calculations we considered the order (μ, γ, δ): bias first, then information, and noise last. If we had computed these contributions in a different order, such as (γ, δ, μ), we could have gotten different contributions, and it is unclear which order should be preferred.

Fortunately, there is a related problem in cooperative game theory that analyzes the individual contributions of team members. We use the solution to this problem here by treating bias, noise, and information as three "team members" working together to improve accuracy. In cooperative game theory, the common solution is to take the average of all possible orders (in our case, six). The averages, known as Shapley values (Shapley 1953), become the component-specific contributions to the overall change ΔBrS. These values have many desirable properties (for a review see Hart 1989) including the fact that they sum to the overall change ΔBrS. Note that although our focus is on Brier scores, we can calculate individual contributions in this manner for any performance metric such as logarithmic or spherical scoring rules (e.g., Gneiting and Raftery 2007).

The following stylized example illustrates why order matters and how individual contributions are calculated.

EXAMPLE 2. Suppose $\mu^* = 0$ so that the base rate is 0.5. Consider calculating the Shapley values for the following groups of forecasters:

$$\text{Control Group:} \quad \begin{cases} \text{Bias:} & \mu_0 = 0 \\ \text{Information:} & \gamma_0 = 0 \\ \text{Noise:} & \delta_0 = \infty \end{cases}$$

$$\text{Treatment Group:} \quad \begin{cases} \text{Bias:} & \mu_0 = 0 \\ \text{Information:} & \gamma_0 = 1 \\ \text{Noise:} & \delta_0 = 0 \end{cases}$$

Both groups are unbiased: $\mu_0 = \mu_1 = 0$. The control group has no information ($\gamma_0 = 0$) and an extremely high level of noise ($\delta_0 = \infty$). Therefore, purely due to noise, the predictions of this group oscillate between 0.0 and 1.0 with equal probability and independent of the actual outcome. By chance, predictions match the outcome half of the time and the other half of the time the group predicts the opposite (0.0 when the outcome is 1, and vice versa). The expected Brier score then is $0.5 \times 0.0 + 0.5 \times 1.0 = 0.5$. By contrast, the treatment group is perfectly informed ($\gamma_1 = 1$) and has no noise ($\delta_1 = 0$). Predictions are perfect, so the Brier score is always 0.0.

Given that biases are the same for the groups, the order-specific contributions of bias are zero under all orders of parameters. Differences in individual contributions then come from noise and information, and there are two different orders: (δ, γ) and (γ, δ).

Consider (γ, δ) first:

i. Change γ_0: Even if we change the control group's information parameter $\gamma_0 \to \gamma_1 = 1$, the extreme noise continues to dominate and the expected Brier score remains at 0.5.

ii. Change $\delta_0|\gamma_0 = 1$: Conditional on full information $\gamma_0 = \gamma_1 = 1$, changing the control group's noise parameter $\delta_0 \to \delta_1 = 0$ makes it unbiased ($\mu_0 = 0$), noise-free ($\delta_0 = 0$), and fully informed ($\gamma_0 = 1$). The control group now predicts the outcome perfectly and has a Brier score equal to 0.

To summarize, first changing $\gamma_0 \to \gamma_1 = 1$ has no effect and the Brier score remains at 0.5 but, conditional on $\gamma_0 = \gamma_1 = 1$, changing $\delta_0 \to \delta_1 = 0$ decreases the Brier score from 0.5 to 0. The specific contributions of order due to information and noise then are 0.0 (from 0.5 to 0.5) and 0.5 (from 0.5 to 0.0), respectively.

Consider now the other order, (δ, γ):

i. Change δ_0: If we set the control group's noise parameter $\delta_0 \to \delta_1 = 0$, the control group becomes unbiased ($\mu_0 = 0$) and noise-free ($\delta_0 = 0$) but still remains entirely uninformed ($\gamma_0 = 0$). Therefore no variability remains in their predictions which always equal the base rate of 0.5, yielding a Brier score of $0.5 \times (0.5 - 0)^2 + 0.5 \times (0.5 - 1.0)^2 = 0.25$.

ii. Change $\gamma_0|\delta_0 = 0$: Conditional on no noise $\delta_0 = \delta_1 = 0$, changing the control group's information parameter $\gamma_0 \to \gamma_1 = 1$ makes it unbiased ($\mu_0 = 0$), noise-free ($\delta_0 = 0$), and fully informed ($\gamma_0 = 1$). The control group now predicts the outcome perfectly and the Brier score decreases from 0.25 to 0.0.

To summarize, first changing $\delta_0 \to \delta_1 = 0$ decreases the Brier score from 0.5 to 0.25 and, conditional on $\delta_0 = \delta_1 = 0$, changing $\gamma_0 \to \gamma_1 = 1$ decreases the Brier score from 0.25 to 0.0. The specific contributions of order due to noise and information then are 0.25 (from 0.5 to 0.25) and 0.25 (from 0.25 to 0.0), respectively.

Averaging the order-specific contributions produces overall contributions due to bias, information, and noise, namely Shapley values of 0.0 for bias, 1/8 for information (from $(0 + 0.25)/2$), and 3/8 for noise (from $(0.5 + 0.25)/2$). For this treatment group, reducing noise contributes three times more to accuracy than does increasing information, which underscores the value of noise reduction. Even if a forecaster has perfect information, enough noise can mask it, and the result will be poor accuracy.

3.3. Demonstrating Robustness on Simulated Data

Before applying the BIN model to GJP data, we evaluate it on synthetic data with known parameter values. Ideally, we would perform an exhaustive evaluation over all possible combinations of values of our model parameters. However, our model has 10 parameters that can take a continuum of values. Furthermore, the size of groups varies across questions, and the number of questions varies

across contexts. An exhaustive evaluation cannot be computed or presented succinctly. Choices had to be made.

To simplify, we constructed a controlled synthetic environment that resembles our application to GJP data. We chose parameter values based on estimates of GJP data. First, we fix the base rate of change at 0.21. Second, suppose each comparison group has 75 forecasters, approximately the median number of forecasters in each condition of the GJP data. Third, consider a control group that resembles the untrained individuals. Specifically, let $\mu_0 = 0.50$, $\gamma_0 = 0.20$, and $\delta_0 = 1.00$. We then vary the parameters of the treatment group and report the accuracy of our estimation. Fourth, the covariance in the forecasters' interpreted signals was set to 0.05. They are thus mildly correlated both within and between groups.

We begin with a particular treatment group: $\mu_1 = 0.25$, $\gamma_1 = 0.15$, and $\delta_1 = 0.5$. Therefore this hypothetical treatment reduces bias, noise and information. Figure 2 shows the 95% credible intervals for bias, noise, and information as the number of questions increases from 10 to 200, adding 10 questions each time and re-estimating parameters. Horizontal dashed lines indicate true parameter values. Overall, 95% credible intervals become narrower and gravitate toward the true values as the number of questions increases. Figure 2 illustrates the consistency of Bayesian estimation techniques. In the Good Judgment Project the number of questions in the comparisons of treatment and control groups ranges from 87 to 191. The simulation suggests we will have reasonably good estimates with these datasets.

Although this is only a single simulated dataset, we can obtain a more comprehensive view to the performance of our estimation procedure by repeating the analysis on many synthetic datasets and treatment groups. To do this, we fix the number of questions to 100 and generate 100 datasets with varying combinations of $\mu_1 \in \{-0.5, 0, 0.5\}$, $\gamma_1 \in [0.01, 0.3]$, and $\delta_1 \in [0, 2]$. For each dataset, we then calculate the posterior means of the parameters and their root-mean-squared-errors (RMSE) in estimating the true parameter values.

Figure 3 presents the average RMSE for $\mu_1 = 0.0$ and different pairs of γ_1 and δ_1 in a 2-by-3 grid of panels. The other two values for the bias term, namely for $\mu_1 = -0.5$ and for $\mu_1 = 0.5$, gave very similar results and are discussed in the supplementary material. The top row represents the control group; the bottom row, the treatment group. Values of information and noise in the treatment group, γ_1 and δ_1, range over the x- and y-axes of the panels, respectively. Colors within the plots show the RMSEs, as indicated by the legend on the right side of the figure. Some values are missing because certain parameter configurations are not possible. Under some parameter configurations the multivariate normal distribution of the forecasters' interpreted signals and the outcome-determining variable (see Appendix A for details) is degenerate and hence not feasible.

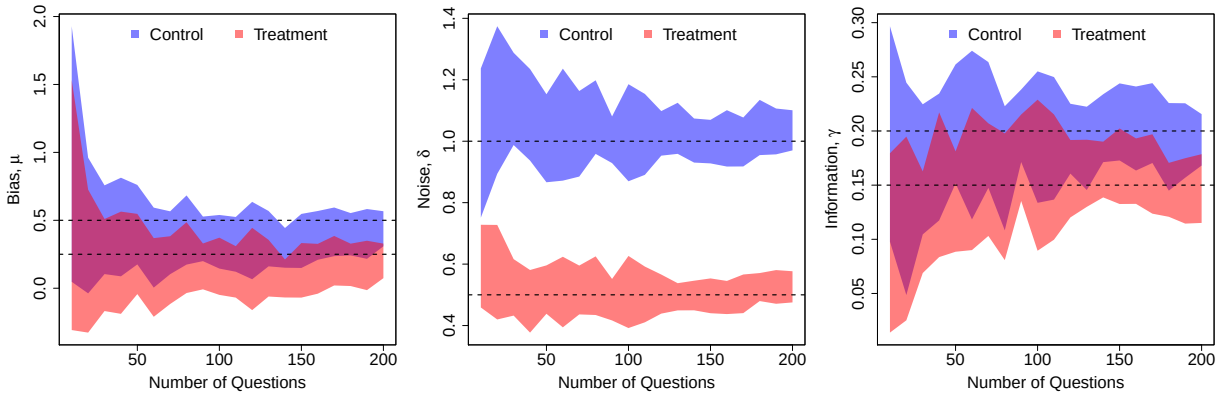

Figure 2 The shaded regions are the 95% credible intervals under varying numbers of questions. Blue represents the control group and red represents the treatment group. Dashed horizontal lines are the true parameter values.

Overall, the RMSE is always below 0.13, and in most cases, well below 0.05, suggesting that our estimation method can capture true values accurately across a wide range of parameter values with the quantity of data in the GJP dataset. The largest RMSEs occur in estimating bias. Recall, however, that bias, noise, and information are on different scales. Bias can take on any value between negative and positive infinity, noise must be non-negative, and information must always be between zero and one. The RMSE in estimating δ_1 increases as δ_1 gets close to 2.0. To illustrate why this happens, consider a forecaster with no bias or information. Suppose that the base rate is 0.5 and the level of noise is very high, say, $\delta_1 = 100$ such that the probability predictions are essentially 0.0 or 1.0 with equal probability. Increasing the level of noise to $\delta_1 = 101$ will have virtually no effect on the distribution of the probability predictions; they will still equal 0.0 or 1.0 with equal probability. As a result, it is difficult to detect the change in δ_1 based on the probability predictions. This explains why higher levels of δ_1 are more difficult to estimate. In practice, however, this is unlikely to be a problem because $\delta_1 = 2$ represents an usually high level of noise. In the GJP dataset the estimated levels of noise ranged between 0.28 and 1.09.

3.4. 30 Days to Resolution: Glossary

Tables 1-3 present results for predictions 30 days prior to outcome resolution as well as pairwise comparisons of control and treatment groups in each column. Table 1 focuses on model parameters across forecaster groups. The sections of Table 1 are:

- *Parameter estimates*: We show posterior means of the parameters of interest and their differences. Under each mean is the 95% credible interval that represents the interval in which the true parameter value falls with 95% probability. Contrast the credible interval with the classical 95% confidence interval that, over many replications of the study, contains the true parameter value 95% of the time. Given that the confidence interval treats parameter values as fixed, probabilistic

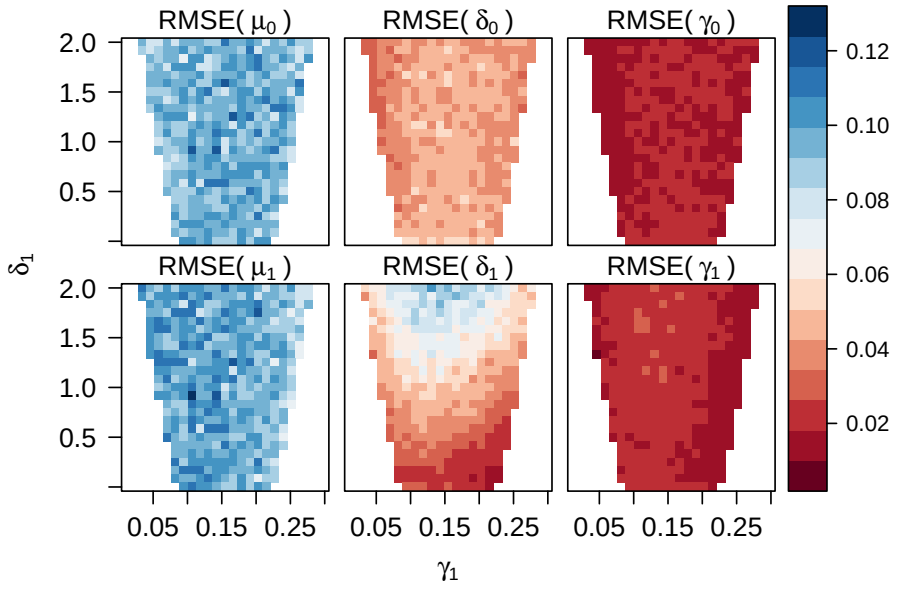

Figure 3 Root Mean Squared Errors (RMSE) in estimation of parameter values of treatment groups. The columns represent bias, noise, and information from left to right, respectively. The top row represents the control group; the bottom row, the treatment group. The true bias in the treatment group is $\mu_1 = 0$. Other choices yield similar results.

statements, such as the interval containing the true parameter value with 95% probability, are not meaningful. But within a Bayesian statistical framework, they are meaningful, and this is how credible intervals should be interpreted.

- *Posterior inferences*: This section provides the posterior probabilities of different events. Compared to the control group, does the treatment group have: (i) less bias, (ii) more information, and (iii) less noise? Intuitively, one can think of these probabilities as the Bayesian analogs of the p-values in classical hypothesis testing – the closer the probability is to 1.0, the stronger the evidence for the hypothesis.

Table 2 links forecasters' ability to each component. The sections of Table 2 are:

- *Predictive performance*: Brier scores of the control and treatment groups.
- *Value of the contribution*: Individual contributions for each treatment.
- *Percentage of control group Brier score*: Individual contributions divided by the expected Brier score of the control group. These values show, in percentage terms, how the change in the Brier score can be attributed to each component.
- *Maximum achievable contribution*: Transformed contributions for a hypothetical treatment that induces perfect accuracy (no bias, no noise, full information). These values can be seen as theoretical limits on improvement for a given component (bias, information or noise).

Table 3 describes the data used in each comparison:

- *Data summary*: Number of questions and the median number of treatment and control group predictions per question. The number of questions varies across conditions because some treatment conditions were present in all four years of the tournament, whereas others were only present in one or two. Furthermore, not every forecaster predicted every question. In rare cases in which no member of either group predicted a question, that question was dropped.

3.5. 30 Days to Resolution: Discussion

Posterior estimates in Table 1 show that the base rate of change is around $\Phi(\mu^*) \approx \Phi(-0.88) = 0.189$ and that all groups, except that with trained and super-team conditions, had upward bias. Groups exaggerated change: they assigned excessive probabilities to changes in the status quo. Second, the amount of information varied across groups (from 0.21 to 0.60). (Recall that an information level of 1.0 represents full information). Super teams were more informed than other groups. Third, all groups had noise but some had much more than others.

Super teams were the most informed, least noisy, and least biased. Given that superforecasters were selected based on their previous excellent performance, their information, bias, and noise can been seen as approximations of what is achievable within this forecasting environment.

Posterior inferences show whether the treatment group was significantly better than the control group on each component. First, all treatments reduced noise with an estimated posterior probability of virtually 1.0[2]. Second, all treatments, except probability training in individuals, increased information with posterior probabilities ranging from 0.97 to 1.00. Third, only teaming reduced bias, with posterior probabilities ranging from 0.99 to 1.0. In the other cases, we expected bias reductions, but posterior probabilities for bias ranged from only 0.72 to 0.91. Even though our methodology can be used to compare the magnitude of positive or negative bias separately across the two groups, we focus, for simplicity, only on the absolute value of the bias: $|\mu_0|$ and $|\mu_1|$. So, a forecaster with, e.g., $\mu_0 = -0.1$ is more biased than one with $\mu_1 = 0.05$.

These results suggest that reducing noise is easier than reducing bias. Noise could be tamped down via training, teaming, or tracking, whereas bias could only be reduced via a particular type of teaming, a manipulation that encouraged forecasters to be actively open-minded and to grapple with dissonant arguments.

One reason why the interventions reduced noise is that training improved forecasters' understanding of probability by giving them a more granular understanding of uncertainty that helped

[2] Estimated probabilities of 0.0 or 1.0 are possible because the calculation is based on a finite posterior sample of $4,000$ draws of which the first $2,000$ were used for burn-in. The probabilities then represent the proportion of the final $2,000$ parameter draws in which the treatment group is superior to the control group. The rounding error could be reduced by computing a large posterior sample. The results, however, would not be qualitatively different. For instance, an estimated posterior probability of 0.0 could become, say, 0.00001, leading to the same conclusions.

them translate their verbal hunches into a probability metric, like "distinct possibility" and "very likely," that carry a wide range of quantitative meanings (Friedman et al. 2018). Insofar as trained individuals more reliably recognize that a forecast of 0.95 implies $19:1$ betting odds whereas 0.9 implies much lower odds, $9:1$, we should expect their judgments to be more consistent – and less noisy.

Of course, it is not surprising that forecasters who invest more cognitive effort in forming their predictions, are likelier to report probabilities that reflect their true beliefs (information) about the event. But there are different ways to encourage cognitive effort and different pathways via which greater effort can translate into accuracy. For instance, teaming is one motivator. Forecasters working in teams can see each other's predictions, which introduces a social aspect to forecasting. Extreme predictions in the wrong direction can cause a forecaster to lose status. As a result, individuals are likely to become less erratic and more circumspect in their predictions, thus reducing their noise.

Tracking selects "superforecasters" based on their capacity to deliver consistently lower Brier scores over time and across topics. Given that it is logically impossible for noisy forecasters to qualify as superforecasters, it is not be surprising to find less noise in their judgments.

Reducing bias, however, may be harder than reducing noise due to the tenacious nature of certain cognitive biases. Kahneman (2011) illustrates this point with the Müller-Lyer illusion that shows two lines, one with arrow heads pointing inwards and the other with arrow heads pointing outwards. The line with inward-pointing arrows looks longer even though the lines are equally long. Interestingly, however, even after we measure the lines with a ruler, we cannot help seeing one line as longer. If cognitive biases are indeed as stubborn as perceptual illusions, we should expect biases to persist even after they have been pointed out to the forecasters.

Our results, however, suggest that not all biases are as irresistible as perceptual illusions (Arkes 1991, Lerner and Tetlock 1999). First, superforecasters showed relatively little bias, suggesting that close-to-unbiased forecasting is humanly achievable in our forecasting environment. Second, teaming reduced bias, which may well reflect how teams were instructed to interact, second-guessing each other to avoid excessive conformity/herding (Tetlock and Gardner 2016). Consistent with past work, properly organized and incentivized groups can check some forms of cognitive biases (e.g., Kerr et al. 1996).

Although the posterior probabilities provide evidence that the interventions decreased bias and noise and increased information, these probabilities do not tell us whether the changes are large enough to be important. A treatment can reduce bias significantly but the decrease could be so small that it may have little real-world relevance. To understand the practical relevance of each change, the *Percentage-of-control-group Brier scores* in Table 2 provides the normalized individual

contributions, as described in Section 3.4. These data reaffirm our claim that noise reduction is most important to improving accuracy. Training improves accuracy almost entirely through noise. Both teaming and tracking improve accuracy via all three components: in order of importance, noise, bias and information.

The *Maximum achievable contributions* in Table 2 should interest any organization that employs forecasters and cares about accuracy. This section of the table shows that forecasters fall short of perfect forecasting due more to noise than bias or lack of information. Eliminating noise would reduce the Brier score of the control group by roughly 50%; eliminating bias would yield a roughly 25% cut; increasing information would account for the remaining 25%. In sum, from a variety of analytical angles, reducing noise is roughly twice as effective as reducing bias or increasing information.

3.6. Other Horizons: Discussion

Unfortunately, space constraints prevent us from showing Tables 1-3 for all time horizons. Therefore, to examine how our findings vary across time horizons and to gauge how bias, noise, and information evolve over time, we present the relative contributions of the difference between Brier scores of the Control Group vs Treatment Group across Days 1 to 60.

Figure 4 shows differences in the contributions to Brier scores for examples of training, teaming and tracking. Additional comparisons are presented in the supplementary material. For any given day, colored bars represent the difference in the expected Brier scores due to noise (red), information (green), and bias (blue). The sum of the three sub-bars is the difference in Brier scores due to the intervention. Note that all changes decrease from the right (Day 60) to the left (Day 1).

The left panel reveals that the primary reason trained teams are better than untrained teams is noise reduction. Red bars are larger than blue or green bars across the horizon. Red bars are larger at the longer horizons (around 30-60 days) than the shorter ones (1-30 days). Bias reduction (blue bars) contributes to the better performance of trained teams at longer horizons. The relative contribution of information (green bars) does not change much over time or play an important role in Brier score differences. In sum, training improves accuracy almost entirely through noise reduction and, to a lesser extent, bias reduction.

The center panel shows the effects of teaming on forecasters who are trained. Again, the primary reason teams are superior to individuals is noise reduction (red bars). Noise reduction contributes almost half of the difference in Brier scores, and this relative contribution remains constant over time. Bias and information share the remaining portions and evolve in opposing manners. Teams reduce bias more at longer horizons and boost information more at shorter horizons. As time progresses, more information becomes available, and teaming – unlike training in the left panel – allows forecasters to harness the information.

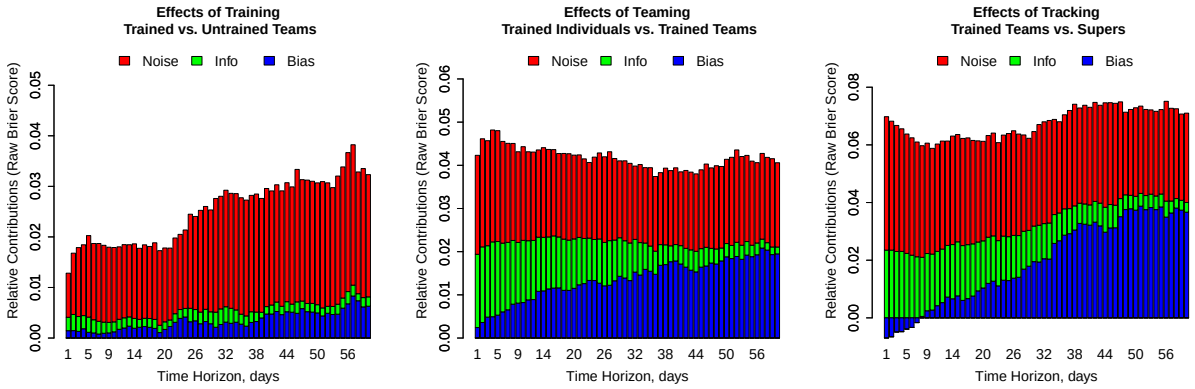

Figure 4 Contributions of noise, information, and bias in reducing Brier scores at varying time horizons.

The right panel presents the tracking of elite teams (superforecasters) compared to trained regular teams. Relative contributions exhibit a pattern similar to that of teaming in the center panel. At Day 60, elite teams have less noise and less bias. Information differences are not great. As the resolution date draws nearer, elite teams have less noise, more information, and slightly more bias than regular teams. There is a negative contribution of bias for superforecasters around 10 days before the resolution date. This means that elite teams have more bias than regular teams. A close inspection of the data reveals that both averages begin above the base rate but, whereas the average forecast among trained teams does not drop below the base rate, the average forecast of the "superforecasters" does. Eventually, as the horizon shrinks (less than 10 days), "superforecasters"" have a bias that is larger than that of trained teams. As a result, the contribution of the difference in Brier scores appears to be negative. One explanation is that top performers tried too hard to win by making forecasts of 0.0 even though there was still a week remaining for surprises. This is a response bias (as opposed to a perceptual bias) that is presumably the result of incentives in the tournament.

4. General Discussion

Of the three logical paths to improving accuracy – tamping down noise and bias or ramping up signal detection – noise reduction consistently emerged as the most important driver of the three experimental interventions: training forecasters in basic principles of probabilistic reasoning, assembling forecasters into open-minded teams, and tracking the best forecasters each year into elite teams. Whenever an intervention boosted accuracy, it worked mainly by suppressing random errors in judgment. Curiously, the original intent of the training intervention was to reduce bias by encouraging forecasters to adopt the "outside view" (Kahneman 2011) and use base rates. The intent of teaming was to prevent biases such as groupthink and share useful information. Lastly, the rationale for tracking was to assemble the most skillful signal detectors into teams in which they could combine their insights and reduce bias and noise and increase information.

These results are striking and raise many questions. For instance, if noise reduction was the dominant, largely unanticipated, driver of the benefits conferred by the experimental interventions, what are the drivers of noise reduction? Can we isolate which facets of these somewhat multifaceted interventions tamped down the noise in human judgment? In retrospect, it makes sense that debiasing training, which stresses the value of anchoring initial probability judgments in base rates – would have the net effect of stabilizing forecasts. It makes sense that teaming would make it easier for forecasters to converge on the more reliable crowd judgment. And it makes sense that tracking the best forecasters into elite teams would make it easier still for those forecasters to converge on more reliable crowd aggregates. These were not however the process stories originally emphasized in the publications from the Good Judgment Project, which put more weight on the capacity of training to tamp down bias (Chang et al. 2016) and of "superforecasters" to be more skillful signal detectors (Mellers et al. 2015).

It is worth stressing the practical and theoretical significance of two discoveries that the BIN model has revealed about top performers in forecasting tournaments. First, the current results suggest an important qualification on past characterizations of top performers ("superforecasters") in both the research literature (Mellers et al. 2015) and in more popular outlets (Tetlock and Gardner 2016). The emphasis in earlier work was on the insightful ways in which top performers extracted subtle predictive clues that others missed or avoided being gulled by pseudo-diagnostic cues that others were misled into using. Our data do not negate the accomplishments, but they do cast the data in a different light. "Superforecasters" may owe their success more to superior discipline in tamping down measurement error, than to incisive readings of the news that others cannot replicate. They generated low-variance, low-Brier score forecasts.

Second, tournaments may over-incentivize top performers, driving them to do things that are suboptimal from a Brier-scoring perspective, such as excessive extremizing as question-closing dates loom. Just because tournament designers made Brier-score minimization the official goal does not mean that human players internalized that imperative. Top performers may want, above all, to "come in first" and that may tempt them to make over-confident claims. For instance, the decision calculus for top performers may take this form: "If I adjust my forecast to a probability of 0.005 even when I truly believe the probability is 0.05, it will yield only a tiny benefit in my Brier score (0.000025 vs 0.0025) but it will do so on the vast majority of 0.05 probability events that do not materialize (99% of the events, for which the supers' average probability forecast five days before the resolution were less than 0.05, did not occur). Those tiny advantages will accumulate and potentially put me in first place. Of course, I realize there would be a steep scoring penalty when these almost-slamdunk events actually materialize (the remaining 1% of events). And that would degrade my accuracy score. I might even fall into the lower ranks of the top performers.

But the risk is worth taking for a shot at being the best of the best." By contrast, trained teams were more cautious in their late predictions: none of the events, for which their average probability forecast five days before the resolution was less than 0.05, occurred. We cannot say for sure that this decision calculus drives the bias bump that the BIN model discovered but we see it as the most parsimonious explanation for the bias-blip among "superforecasters" toward the end of forecasting time horizons.

There is much to learn about noise as a source of judgmental failures and as a driver of successful interventions. We see value in distinguishing four pathways by which noise reduction can be achieved: a) disciplining the internal judgment processes of forecasters by, for instance, required employee participation in noise audits (Kahneman et al. 2016) or in training exercises (Chang et al. 2016) aggregating the judgments of forecasters either through institutional interventions such as prediction markets (Wolfers and Zitzewitz 2004) or through purely statistical means (Larrick and Soll 2006, Budescu and Chen 2014, Satopää et al. 2014, Prelec et al. 2017); c) interventions aimed at simplifying the external world by, for instance, filtering out misleading or low-diagnosticity sources in the news environment and lightening the cognitive load on forecasters (Lazer et al. 2018); and d) the most radical of the measures, replacing human judges with machine learning algorithms, as has been done quite successfully in a number of domains (e.g., targeting restaurants for health inspections (Kang et al. 2013), identifying teenagers at highest risk for committing crimes (Chandler et al. 2011) and deciding which defendants should be held awaiting adjudication of their cases (Kleinberg et al. 2017).

We should, of course, beware of the potential costs of these noise reduction strategies, including the risks of inducing too much uniformity in how forecasters interpret environmental signals or how slowly forecasters respond to environmental changes. That said, however, the current data are suggestive. Noise may well be the easiest source of error to correct – and organizations looking for cost-effective methods of boosting forecasting accuracy should give noise-reduction tools and training serious consideration.

Finally, we close by addressing a standard objection often raised to assigning probability estimates to the ostensibly "unique" or "one-off" in the IARPA tournaments (Tetlock 2017). Let's assume, provisionally, these skeptics are right that it is impossible to separate bias, information and noise in predictions of singular events. After all, putting aside situations in which forecasters declare an event impossible ($p = 0.0$) and the event occurs or declare an event a sure thing ($p = 1.0$) and it does not occur, all other probability assessments of singular events are indeed indeterminate. For instance, we will never know whether Nate Silver's 70% prediction of a Hillary Clinton victory in 2016 was accurate or inaccurate due to bias or noise (Kennedy et al. 2018). We cannot rerun history and measure how closely forecasters' estimated distributions of possible worlds correspond

to actual distributions. But the "singular-event" objection does not apply. For we are applying the BIN model to predictions of multiple "singular-events," or to be precise, to a set of ostensibly singular events within which it is logically possible to separate bias, information and noise as drivers of forecasting accuracy. If probability judgments of these events were as meaningless as the standard objection implies, it should not have been possible to identify systematic individual difference correlates of accuracy or experimental interventions that boost accuracy. In this sense, the BIN model shows that the standard objection over-reaches and needs reformulation.

One promising direction for future work is to extend the BIN model into simulated worlds in which researchers can rerun history and it becomes feasible to make probability claims about singular events, not just aggregated sets of events. In these worlds, researchers can compare forecasters' probability distributions of possible worlds against "actual" distributions – in effect, exercises in counterfactual forecasting. Just as it is possible in real-world tournaments to identify better forecasters and better forecasting methods, it will be possible to do the same in simulated worlds – and more. Researchers will then be well-positioned to separate psychological noise (the focus of the current work) from environmental noise (randomness inherent in causal processes generating outcomes). Good forecasters should be more attuned to when reruns of history, from various starting points, yield close-to-deterministic outcomes (98% of the time this business fails) and highly stochastic outcomes (a wide distribution of possible outcomes, from bankruptcy to billion-dollar valuations). And it will be an open question whether forecasting skills in simulated worlds will transfer back to the real world.

Appendix. Technical Details

A. Model Implementation

Suppose there are K outcomes. Denote the kth outcome with $Y_k \in \{0,1\}$, where $Y_k = 1$ if the event happens and $Y_k = 0$ otherwise. This is determined by a normal random variable Z_k^* such that $Y_k = \mathbf{1}(Z_k^* > 0)$ for all $k = 1, \ldots, K$. Suppose there are $N_{0,k}$ and $N_{1,k}$ treated and control forecasters, respectively, making predictions for the kth outcome. Collect their normally distributed interpretations into vectors $\mathbf{Z}_{0,k} = (Z_{0,1} \ldots Z_{0,N_{0,k}})'$ and $\mathbf{Z}_{1,k} = (Z_{1,1} \ldots Z_{1,N_{1,k}})'$. Then, Assumptions 1 and 2 give

$$\begin{pmatrix} Z_k^* \\ \mathbf{Z}_{0,k} \\ \mathbf{Z}_{1,k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu^* \\ (\mu^* + \mu_0)\mathbf{1}_{N_{0,k}} \\ (\mu^* + \mu_1)\mathbf{1}_{N_{1,k}} \end{pmatrix}, \begin{pmatrix} 1 & \boldsymbol{\Sigma}_{0,k}' & \boldsymbol{\Sigma}_{1,k}' \\ \boldsymbol{\Sigma}_{0,k} & \boldsymbol{\Sigma}_{00,k} & \boldsymbol{\Sigma}_{01,k} \\ \boldsymbol{\Sigma}_{1,k} & \boldsymbol{\Sigma}_{01,k}' & \boldsymbol{\Sigma}_{11,k} \end{pmatrix} \right), \tag{3}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{gg,k} &= \mathrm{Cov}(\mathbf{Z}_{g,k}, \mathbf{Z}_{g,k}) = \mathbf{I}_{N_{g,k}}(\delta_g + \gamma_g - \rho_g) + \boldsymbol{J}_{N_{g,k} \times N_{g,k}} \rho_g \text{ for } g \in \{0,1\}, \\ \boldsymbol{\Sigma}_{g,k} &= \mathrm{Cov}(\mathbf{Z}_{g,k}, Z_k^*) = \mathbf{1}_{N_{g,k}} \gamma_g \text{ for } g \in \{0,1\}, \\ \boldsymbol{\Sigma}_{01,k} &= \mathrm{Cov}(\mathbf{Z}_{0,k}, \mathbf{Z}_{1,k}) = \boldsymbol{J}_{N_{0,k} \times N_{1,k}} \rho_{01}, \end{aligned}$$

$$Y_k = \mathbf{1}(Z_k^* > 0),$$
$$p_{g,j} = \Phi\left(\frac{Z_{g,j}}{\sqrt{1-\gamma_g}}\right) \text{ for } j = 1,\ldots,N_{g,k} \text{ and } g \in \{0,1\},$$

$\boldsymbol{J}_{a\times a}$ is a $a \times a$ matrix of ones, $\mathbf{I}_a$ is a $a \times a$ identity matrix, and $\mathbf{1}_a$ is a vector of a ones. The additional parameters ρ_{01}, ρ_1, and ρ_0 describe the covariances of the interpretations across and within each group. Given that these covariances are not directly linked to the outcome, they can stem from shared information or noise. These parameters are not of direct interest in our application but must be included for the sake of model completeness.

Equation (3) gives the likelihood for the kth event and its predictions. The joint likelihood of the K events is constructed from (3) by assuming the predictions and outcomes to be (conditional on the model parameters) independent and identically distributed across different events. More specifically, Assumption 3 gives

$$f(\mathbf{Z}^*, \mathbf{Z}_0, \mathbf{Z}_1|\boldsymbol{\theta}) = \prod_{k=1}^{K} f(Z_k^*, \mathbf{Z}_{0,k}, \mathbf{Z}_{1,k}|\boldsymbol{\theta}), \tag{4}$$

where $\boldsymbol{\theta} = (\mu^* \ \mu_0 \ \mu_1 \ \gamma_0 \ \gamma_1 \ \delta_0 \ \delta_1 \ \rho_0 \ \rho_1 \ \rho_{01})'$, $\mathbf{Z}^* = (Z_1^* \ldots Z_K^*)'$, $\mathbf{Z}_0 = (\mathbf{Z}_{0,1}' \ldots \mathbf{Z}_{0,K}')'$, and $\mathbf{Z}_1 = (\mathbf{Z}_{1,1}' \ldots \mathbf{Z}_{1,K}')'$, and $f(Z_k^*, \mathbf{Z}_{0,k}, \mathbf{Z}_{1,k}|\boldsymbol{\theta})$ is the likelihood of (3).

To compute the posterior distribution of the parameters, we use the likelihood (4) together with a flat, non-informaive prior distribution on $\boldsymbol{\theta}$; that is, $\pi(\boldsymbol{\theta}) \propto 1$. The parameters are then estimated with a Markov Chain Monte Carlo (MCMC) technique called Hamiltonian Monte Carlo sampling. This is a standard estimation procedure in Bayesian statistics which we implement using Stan software package (Carpenter et al. 2017). The final output is a sample from the joint posterior distribution of the parameters (see, e.g., Gelman et al. 2013 for a description of Bayesian estimation techniques). Joint estimation allows us to make statistical significance statements about parameters across the two groups of forecasters.

B. Analytical Expression for the Expected Brier Score

PROPOSITION 1. *The value of the expected Brier score is*

$$BrS(\mu_g, \delta_g, \gamma_g, \mu^*) = \Phi(\mu^*) - 2\Phi_2(\boldsymbol{\mu}' \mid \boldsymbol{\Omega}') + \Phi_2(\boldsymbol{\mu}'' \mid \boldsymbol{\Omega}''),$$

where

1. $\Phi(\cdot)$ *is the standard Gaussian CDF;*
2. $\Phi_2(\cdot \mid \boldsymbol{\Omega})$ *is the bivariate Gaussian CDF with zero mean vector and covariance matrix* $\boldsymbol{\Omega}$*;*
3. *the vectors* $\boldsymbol{\mu}' = \left[\mu^* \ \frac{\mu^*+\mu_g}{\sqrt{1-\gamma_g}}\right]$ *and* $\boldsymbol{\mu}'' = \left[\frac{\mu^*+\mu_g}{\sqrt{1-\gamma_g}} \ \frac{\mu^*+\mu_g}{\sqrt{1-\gamma_g}}\right]$*; and*
4. *the matrices* $\boldsymbol{\Omega}' = \begin{bmatrix} 1 & \frac{\gamma_g}{\sqrt{1-\gamma_g}} \\ \frac{\gamma_g}{\sqrt{1-\gamma_g}} & \frac{1+\delta_g}{1-\gamma_g} \end{bmatrix}$ *and* $\boldsymbol{\Omega}'' = \begin{bmatrix} \frac{1+\delta_g}{1-\gamma_g} & \frac{\gamma_g+\delta_g}{1-\gamma_g} \\ \frac{\gamma_g+\delta_g}{1-\gamma_g} & \frac{1+\delta_g}{1-\gamma_g} \end{bmatrix}$.

Proof of Proposition 1 Consider the definition of the expected Brier score:

$$\begin{aligned}\mathrm{BrS}(\mu_g, \delta_g, \gamma_g, \mu^*) &= \mathbb{E}_{Y,Z_g}\left\{\left[Y - \Phi\left(\frac{Z_g}{\sqrt{1-\gamma_g}}\right)\right]^2\right\} \\ &= \mathbb{E}_{Y,Z_g}\left\{Y^2 - 2\Phi\left(\frac{Z_g}{\sqrt{1-\gamma_g}}\right)Y + \Phi^2\left(\frac{Z_g}{\sqrt{1-\gamma_g}}\right)\right\}.\end{aligned}$$

Given that $Y \in \{0,1\}$, the first term is equal to $\mathbb{E}[Y^2] = \mathbb{E}[Y] = \Phi(\mu^*)$. To compute the second term, introduce a standard normal random variable ε and rewrite the term equivalently as:

$$\mathbb{E}_{Y,Z_g}\left\{\Phi\left(\frac{Z_g}{\sqrt{1-\gamma_g}}\right)Y\right\} = \mathbb{E}_{Z^*,Z_g,\varepsilon}\left\{\mathbb{P}\left(\varepsilon < \frac{Z_g}{\sqrt{1-\gamma_g}}\right)\mathbf{1}(Z^*>0)\right\},$$

which is equal to $\mathbb{E}_{Z^*,Z_g,\varepsilon}\left\{\mathbb{P}\left(\frac{Z_g}{\sqrt{1-\gamma_g}} - \varepsilon > 0\right)\mathbb{P}(Z^*>0)\right\}$. This value is equal to the probability that a bivariate normal random variable given by $\begin{bmatrix} Z^* & Z_g - \varepsilon \end{bmatrix}$ is (coordinate-wise) greater than the zero vector. The mean of this random variable is $\boldsymbol{\mu}'$ and its covariance matrix is $\boldsymbol{\Omega}'$, implying that:

$$\mathbb{E}_{Z^*,Z_g,\varepsilon}\left\{\mathbb{P}\left(\frac{Z_g}{\sqrt{1-\gamma_g}} - \varepsilon > 0\right)\mathbb{P}(Z^*>0)\right\} = \Phi_2(\boldsymbol{\mu}' \mid \boldsymbol{\Omega}').$$

The third term is computed similarly to the second one; instead of introducing just one random variable ε, introduce two indepedent standard Gaussians ε_1 and ε_2. The mean and covariance matrix of the resulting random variable are $\boldsymbol{\mu}''$ and $\boldsymbol{\Omega}''$, respectively. □

References

Arkes HR (1991) Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin* 110(3):486.

Armstrong JS (2001) *Principles of Forecasting: A Handbook for Researchers and Practitioners*, volume 30 (New York: Springer).

Bliss CI (1934) The method of probits. *Science* 79:38–39.

Budescu DV, Chen E (2014) Identifying expertise to extract the wisdom of crowds. *Management Science* 61(2):267–280.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* 76(1):1–32, ISSN 1548-7660, URL `http://dx.doi.org/10.18637/jss.v076.i01`.

Chandler D, Levitt SD, List JA (2011) Predicting and preventing shootings among at-risk youth. *American Economic Review* 101(3):288–92.

Chang W, Chen E, Mellers B, Tetlock P (2016) Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting yournaments. *Judgment and Decision making* 11(5):509.

Friedman JA, Baker JD, Mellers BA, Tetlock PE, Zeckhauser R (2018) The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly* 62(2):410–422.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian Data Analysis* (New York: Chapman and Hall).

Gilovich T, Griffin D, Kahneman D (2002) *Heuristics and biases: The Psychology of Intuitive Judgment* (New York: Cambridge University Press).

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Hart S (1989) Shapley value. Eatwell J, Milgate M, Newman P, eds., *The New Palgrave: Game Theory*, 210–216 (Norton & Company).

Kahneman D (2011) *Thinking, fast and slow* (New York: Farrar, Straus & Giroux).

Kahneman D, Rosenfield AM, Gandhi L, Blaser T (2016) Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* 94(10):38–46.

Kang JS, Kuznetsova P, Luca M, Choi Y (2013) Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443–1448.

Kennedy C, McGeeney K, Keeter S, Patten E, Perrin A, Lee A, Best J (2018) Implications of moving public opinion surveys to a single-frame cell-phone random-digit-dial design. *Public Opinion Quarterly* 82(2):279–299.

Kerr NL, MacCoun RJ, Kramer GP (1996) Bias in judgment: Comparing individuals and groups. *Psychological review* 103(4):687.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1):237–293.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science* 52(1):111–127.

Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, et al. (2018) The science of fake news. *Science* 359(6380):1094–1096.

Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psychological bulletin* 125(2):255.

McCullagh P (2019) *Generalized Linear Models* (Routledge).

Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, Baker J, Hou Y, Horowitz M, et al. (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3):267–281.

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, Moore D, Atanasov P, Swift SA, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25(5):1106–1115.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgements: Eliciting Experts' Probabilities* (Chichester: John Wiley & Sons).

Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532.

Ravishanker N, Dey DK (2001) *A First Course in Linear Model Theory* (London: CRC Press).

Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH (2014) Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting* 30(2):344–356.

Satopää VA, Pemantle R, Ungar LH (2016) Modeling probability forecasts via information diversity. *Journal of the American Statistical Association* 111(516):1623–1633.

Shapley LS (1953) A value for n-person games. *Contributions to the Theory of Games* 2(28):307–317.

Tetlock PE (2017) *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press), 2nd edition.

Tetlock PE, Gardner D (2016) *Superforecasting: The Art and Science of Prediction* (New York: Crown Publishing).

Wolfers J, Zitzewitz E (2004) Prediction markets. *Journal of Economic Perspectives* 18(2):107–126.