Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Bias, Information, Noise: The BIN Model of Forecasting

Ville A. Satopää, Marat Salikhov, Philip E. Tetlock, Barbara Mellers

To cite this article:
Ville A. Satopää, Marat Salikhov, Philip E. Tetlock, Barbara Mellers (2021) Bias, Information, Noise: The BIN Model of Forecasting. Management Science 67(12):7599-7618. https://doi.org/10.1287/mnsc.2020.3882

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 12, December 2021, pp. 7599–7618
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Bias, Information, Noise: The BIN Model of Forecasting

Ville A. Satopää,[a] Marat Salikhov,[b] Philip E. Tetlock,[c] Barbara Mellers[c]

[a] INSEAD, Fontainebleau 77300, France; [b] Yale School of Management, New Haven, Connecticut 06511; [c] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Contact: ville.satopaa@insead.edu, https://orcid.org/0000-0003-4953-2254 (VAS); marat.salikhov@yale.edu, https://orcid.org/0000-0003-0188-5472 (MS); tetlock@wharton.upenn.edu (PET); mellers@wharton.upenn.edu (BM)

Received: November 17, 2019
Revised: July 6, 2020; September 20, 2020
Accepted: October 5, 2020
Published Online in Articles in Advance: February 10, 2021

https://doi.org/10.1287/mnsc.2020.3882

Abstract. 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 in probabilistic reasoning, anti-groupthink teaming, and tracking of talent. Drawing on these data, we propose a Bayesian BIN model (Bias, Information, Noise) for disentangling the underlying processes that enable forecasters and forecasting methods to improve—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 that noise reduction plays a surprisingly consistent role across all three methods of enhancing performance. We see the BIN method as useful in focusing managerial interventions on what works when and why in a wide range of domains. An R-package called BINtools implements our method and is available on the first author's personal website.

History: Accepted by Manel Baucells, decision analysis.
Funding: This work was supported by the Intelligence Advanced Research Projects Activity [Grant 140D0419C0049] and the INSEAD-Wharton Alliance.
Supplemental Material: Data and the supplementary material are available at https://doi.org/10.1287/mnsc.2020.3882.

Keywords: 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 information, as well as data of uncertain or questionable reliability. Forecasters must make their 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, individuals might display systematic tendencies 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, Brighton and Gigerenzer 2015). Bias is systematic, so it should be possible, at least in principle, to identify its direction and magnitude.

Noise is unpredictable, nonsystematic error. By definition, it is impossible to anticipate the direction or magnitude of random errors. Kahneman et al. (2016) argue that the research literature on noise is much less developed than the literature on bias because noise is more difficult for human observers, wired up to detect patterns, to see or even accept. It is easy to construct causal explanations of bias that invoke character flaws in the forecasters—hubris, rigidity, prejudice, favoritism—but noise defies causal explanation or categorization.

Separating noise from bias in probabilistic forecasts of a single event is difficult, arguably impossible. However, if we have access to forecasters' predictions about multiple events, we can disentangle expected levels of noise and bias, and treat their relative magnitudes 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. Our model contributes to the literature on Bayesian methods of cognitive modeling (e.g., Lee 2018) by describing differences between two groups of forecasters, which we denote as control and treatment, and hence by allowing us to carry out useful inference, such as calculating the posterior probabilities of the treatment reducing bias, diminishing noise, or increasing information.

The remainder of the article applies the model to data from a multiyear, 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 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 is previously thought (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. We decompose forecasters' signal-extraction process into three components: bias, information, and noise. Bias refers to 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 underestimation of probabilities. Partial information is the informational value of the subset of signals that forecasters use, relative to full information that would permit forecasters to achieve omniscience. Finally, noise is the residual variability that is independent of the outcome. To illustrate, we start with a simple example.

Example 1. A researcher in a multiround game flips a fair coin twice and forecasters predict the probability of two heads. There are four equally likely outcomes: TT, HT, TH, and HH.

Imagine a forecaster with zero bias and zero noise. Unless the forecaster finds new information, the forecaster should treat the four outcomes as equally likely and predict 0.25, the base rate for the outcome, HH. Suppose now that the forecaster gets new (partial) information, and is told the outcome of the first toss before making a prediction. 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 and produces predictions that are, on average, the base rate of 0.25.

In this example, the forecaster's prediction is equally likely to be 0 or 0.5, so the mean is 0.25, the variance is 0.0625, and the covariance between the predictions and outcomes is 0.0625. The identical variance and covariance is no coincidence. If forecasters are rational Bayesian agents minimizing a proper scoring function, such as the Brier score, the variance in their predictions is entirely driven by partial information (Satopää et al. 2016).

Consider now a more realistic case with bias. In particular, suppose a biased but noise-free forecaster thinks a fair coin is unfair, with the probability of heads being 0.6. A forecaster with no partial information should always predict $0.6 \times 0.6 = 0.36$ for HH, over-shooting the base rate of 0.25. If the forecaster observes the first toss, the forecaster would predict 0.0 or 0.6 for HH, depending on whether the first toss was a T or H, respectively. On average, the forecaster's prediction is now 0.3, which exceeds the base rate.

Lastly, imagine a noisy but bias-free forecaster. The forecaster does not realize that the researcher gave the result of an unrelated coin toss, so, the forecaster predicts 0.0 or 0.5 for HH, depending on whether the unrelated toss is a T or H, respectively. But because this toss is independent of the flips that determine the outcome, variability in the forecaster's prediction does not correlate with the outcome. Suppose now that the researcher tells the forecaster the outcome of two flips, only one of which determines the outcome. The forecaster then predicts 1.0 if the researcher reports HH; otherwise, the forecaster predicts 0.0. The mean of the forecaster's predictions is 0.25 $(0.25 \times 1 + 0.75 \times 0)$ and the variance is 0.1875 $(0.25 \times (0.25 - 1)^2 + 0.75 \times (0.25 - 0)^2)$. The forecaster's average prediction is the base rate, so the forecaster is unbiased. Not all variability in the forecaster's predictions, however, covaries with outcomes. There are eight equally likely cases. In 5/8 of them, the forecaster correctly predicts 0.0; in 1/8, the forecaster correctly predicts 1.0; and in the remaining 2/8 cases the forecaster's prediction is incorrect. The covariance between the forecaster's predictions and the outcomes 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 observed the first toss. In this example, partial information is 0.0625 and noise is 0.125 (0.1875 − 0.0625).

Although the components in the BIN model—bias, noise, and (inevitably partial) information—have an intuitive definition, there are alternative models that could capture their effects. Given that our main goal is to disentangle how an intervention improves forecasting, we imposed the requirement that our model describes two groups of forecasters jointly. Only then can we make significance statements about common mechanisms by which treatments affect accuracy. For instance, we can estimate the probability that the treatment group is 20% more accurate than the control group, and that 30% of the difference is attributable to less bias. Even though our model could be applied to a single group, statistical comparisons of two groups

would not be possible if the groups were analyzed independently. We need a joint probability model to make joint statements about two groups.

Borrowing from the statistical theory of generalized linear models (McCullagh and Nelder 1989) and probit regression (Bliss 1934), 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. We follow Satopää et al. (2016) and posit a signal universe that contains all past and future signals of positive or negative relevance to the event. In this realm, the event happens if and only if the sum of all signals is positive.

Forecasters sample and interpret signals with varying skill and thoroughness. The entire signal universe may not be available at any given time, and forecasters are unlikely to predict the future perfectly because they are missing signals that may become available later. The unavailable signals represent aleatory uncertainty.

In addition to relevant signals, the universe contains irrelevant signals. Forecasters may sample relevant signals (increasing partial information) or irrelevant signals (creating noise). Forecasters may center signals incorrectly (creating bias). We can model the accumulation of these signals with continuous variables that exhibit degrees of bias, partial information, and noise in their forecasts. These variables summarize forecasters' (often noisy and biased) interpretations of how the signals they have observed relate to the outcome.

There are, of course, many possible sources of noise in forecasters' judgments, including the inherent noisiness of memory and inference processes in cognitive systems (Wyart et al. 2012, Kahana et al. 2018), and noisiness in translating private often vague hunches into a probability metric (Budescu et al. 2014, Friedman et al. 2018, Van Der Bles et al. 2020). Our focus is, however, on the effects of noise, not the exact causes. To invoke the old distinction between "lumpers," who group things into broad categories, and "splitters," who divide things into smaller categories, we play the role of lumpers here. Our model defines noise as any random variability (on the probit scale) that does not correlate with the outcome.

There are three benefits to modeling outcomes and forecasters with continuous variables: (a) we can represent 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 the 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 have distinct roles in forecasters' judgments that we can separate.

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.*

Continuous latent variables are often modeled as normally distributed variables (e.g., probit regression in Bliss 1934, factor analysis and item response theory in Everett 2013). We make this assumption for analytic tractability. Furthermore, if each forecaster interprets a large number of independent signals and those signals are not heavy tailed, the central limit theorem justifies the normality assumption.

Turning to the technical details of the BIN model, 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 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 $\mathrm{Var}(Z^*) = 1$ and choosing the mean $\mu^* = \mathbb{E}[Z^*]$ such that $\mathbb{P}(Z^* > 0) = p^*$. We then have:

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

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 assigns the probability $p_0 \in (0,1)$ to the event $\{Z^* > 0\}$ based on a normally distributed variable Z_0 that represents the forecaster's 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, forecasters must work with partial information. For instance, they may closely follow news about British politics, which strengthens their signals, but not enough to predict Brexit with certainty. Following Satopää et al. (2016), we introduce partial information in the BIN model with the parameter $\mathrm{Cov}(Z_0, Z^*) = \gamma_0$. The greater γ_0, the more Z_0 covaries with Z^* and the better the forecaster.

Given that both Z_0 and Z^* are on continuous scales, bias equals the difference between their means. 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^*: $\mathrm{Var}(Z_0) - \mathrm{Cov}(Z^*, Z_0) = \delta_0$.

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

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

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, γ_0, varies from no information ($\gamma_0 = 0$) to perfect information ($\gamma_0 = 1$). The information component is bounded by 1 because it represents covariability with Z^* that only has variance 1. If the expert is unbiased ($\mu_0 = 0$), noise-free ($\delta_0 = 0$), and has perfect information ($\gamma_0 = 1$), then $Z_0 = Z^*$ and $\gamma_0 = \mathrm{Var}(Z_0) = \mathrm{Var}(Z^*) = 1$.

The forecaster reports the probability of the event $\{Y = 1\}$, not the interpretation Z_0, so the model describes this process of converting 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:

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

Thus,

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

where the last step follows from the symmetry of the normal distribution.

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

To identify bias, information, and noise, 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 $\mu_0 = 0$.

In theory, this is likelier to hold when we use proper scoring rules, such as the Brier score, to evaluate forecasting accuracy (Brier 1950, Gneiting and Raftery 2007). Proper scoring rules incentivize honest reporting. Forecasters trying to minimize their Brier score should strive to remove any known biases and noise[1] from their predictions. Assuming a Brier-score minimization mindset, any bias or noise remaining in final predictions should be bias and noise of which the forecaster was unaware.

Plugging in this potential for misbeliefs into (1), the forecaster now predicts:

$$\mathbb{P}[Z^* > 0|Z_0] = \Phi\left(\frac{Z_0}{\sqrt{1 - \gamma_0}}\right).$$

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

We can extend the BIN model to groups of multiple forecasters, which we call control and treatment groups 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 components. Allowing each forecaster to have a different set of parameters would lead to parameter proliferation. For interpretive and tractability reasons, 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. Section 5.3 explores relaxing this symmetry assumption.

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 levels 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. 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 $\mathrm{Cov}(Z_{g,j}, Z^*) = \gamma_g$. A summary of the BIN model is:

$$\text{Outcome}: Y = \mathbf{1}(Z^* > 0)$$

$$\text{Control group}: \begin{cases} \text{Bias}: \\ \quad \mu_0 = \mathbb{E}[Z_{0,j}] - \mathbb{E}[Z^*] \\ \text{Information}: \\ \quad \gamma_0 = \mathrm{Cov}(Z_{0,j}, Z^*) \\ \text{Noise}: \\ \quad \delta_0 = \mathrm{Var}(Z_{0,j}) - \mathrm{Cov}(Z_{0,j}, Z^*) \\ j\text{th forecaster's prediction}: \\ \quad p_{0,j} = \Phi\left(\frac{Z_{0,j}}{\sqrt{1-\gamma_0}}\right) \end{cases}$$

$$\text{Treatment group}: \begin{cases} \text{Bias}: \\ \quad \mu_1 = \mathbb{E}[Z_{1,j}] - \mathbb{E}[Z^*] \\ \text{Information}: \\ \quad \gamma_1 = \mathrm{Cov}(Z_{1,j}, Z^*) \\ \text{Noise}: \\ \quad \delta_1 = \mathrm{Var}(Z_{1,j}) - \mathrm{Cov}(Z_{1,j}, Z^*) \\ j\text{th forecaster's prediction}: \\ \quad p_{1,j} = \Phi\left(\frac{Z_{1,j}}{\sqrt{1-\gamma_1}}\right) \end{cases}$$

Thus far, we have described the BIN model for a single event. We now extend the model to multiple events by assuming that predictions and outcomes are independent and identically distributed across events.

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

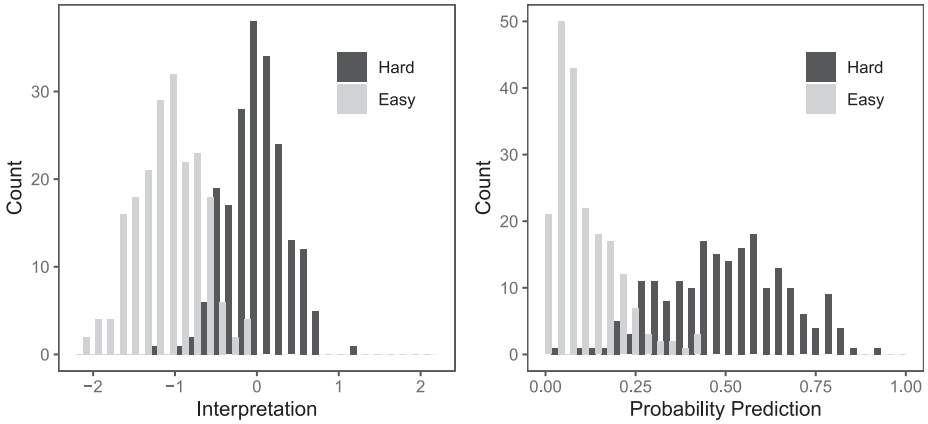
This assumption has two parts. First, the base rate and the forecasters' expected levels of bias, noise, and partial information are the same across events. Forecasters can still have varying interpretations for each event. For easier-to-forecast events, forecasters' interpretations cluster on the correct side of zero, with most predictions pointing in the correct direction. For harder-to-forecast events, the interpretations spread widely around zero, with predictions pointing in opposing directions.

Figure 1 illustrates this with two sets of predictions and interpretations drawn from the BIN model with $p^* = 0.5$, $\mu_0 = 0$, $\gamma_0 = 0.4$, and $\delta_0 = 0$. The gray histogram represents an easier-to-forecast event: all forecasters' interpretations are negative, leading to small probability predictions. Given that this event did not happen, forecasters are accurate. By contrast, the black histogram represents a harder-to-forecast event: the forecasters' interpretations point in different directions, leading to probability predictions that are spread almost symmetrically around 0.5. This event happened, but few forecasters are accurate.

Second, predictions and outcomes are independent across events, conditional on the expected values of bias, information, and noise. Put differently, forecasters' predictions of an outcome provide no additional information about their predictions of other outcomes, as long as we know the base rate and the parameters representing forecasters' bias, noise, and information. This is analogous to the assumption of local independence in item response theory (e.g., Henning 1989).

To illustrate, consider 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

Figure 1. Histograms of Interpretations and Probability Predictions for an Easier-to-Forecast (Gray) and a Harder-to-Forecast (Black) Event



Notes. The event associated with the black histogram occurred, whereas the event associated with the gray histogram did not. Both sets are drawn from a model with the same parameters, illustrating its capacity to capture tasks of widely varying difficulty.

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 is knowledge of their high information and low noise and bias, not of any specific predictions of earlier events.

Taken together, Assumptions 2 and 3 imply that our parameters can be interpreted as average levels of bias, noise, and information within each group of forecasters and across questions. These assumptions help us achieve our goal of understanding both groups' average behavior and making general, not question-specific, statements about bias, information, and noise. Although we cannot make statements at the individual-forecaster level, we can consider different sets of forecasters per event and different numbers of forecasters for each event. Forecasters rarely predict exactly the same events at the same time, so the exchangeability assumption gives us statistical power.

The BIN model strikes a balance between computational tractability and realism. Our assumptions allow us to manage parameter constraints inherent in modeling partial information (see Section 5.3 for details) and to compute the likelihood in constant time (for details, see Section 2 in the supplementary material). No model is an exact description of reality, and violations of the modeling assumptions can distort possible inferences. For instance, empirical data may not follow the normal distribution exactly (violation of Assumption 1), or there can be some degree of dependence across events (violation of Assumption 3). To evaluate the impact of such violations, we subjected our model to sensitivity analyses on simulated data with varying degrees of dependence across outcomes and also on latent variables that were simulated from the (multivariate) t-distribution instead of the multivariate normal distribution. In both cases, our estimation process was not highly sensitive to violations of the assumptions and could recover the true parameters with reasonable accuracy. These results are presented in Section 4 of the supplementary material.

3. Model Estimation

We estimate model parameters with Bayesian statistics that treat parameters as random variables. With any Bayesian model, we need 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 use the Bayes' rule to update the prior distribution in light of the observed data. The updated or posterior distribution describes all uncertainty in the parameters after accounting for the researcher's prior beliefs and data.

Section 2 describes the likelihood function (with technical details in Appendix A). For the prior distribution, we assume a uniform distribution that posits all parameter configurations across the two groups of forecasters to be a priori equally likely. With this assumption, the prior distribution has low impact on the final posterior inference and the data drive the final results. Typically, the posterior distribution cannot be derived analytically. We estimate it 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 of different groups 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 of the likelihood of the treatment group outperforming the control group. Fourth, we calculate how much the treatments improve accuracy via changes in the expected bias, noise, and information.

In our model, the expected Brier score of a forecaster in group $g \in \{0,1\}$ is

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

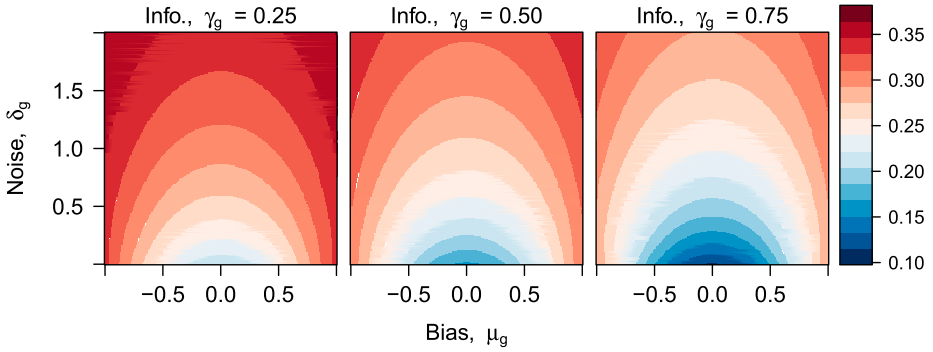
where $Y = \mathbf{1}(Z^* > 0)$, $Z^* \sim \mathcal{N}(\mu^*, 1)$, $Z_g \sim \mathcal{N}(\mu^* + \mu_g, \gamma_g + \delta_g)$, and $\mathrm{Cov}(Z_g, Z^*) = \gamma_g$. We present the analytical expression of (2) and its derivation in Appendix B. To illustrate how the expected Brier score behaves as a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

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

Intuitively, the individual contribution of each parameter could be isolated by changing that parameter

Figure 2. (Color online) Expected Brier Score Under Different Values of Bias, Information, and Noise



Note. The base rate is fixed at $p^* = 0.5$.

and observing the change in the expected Brier score. Those contributions would be:

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

The problem, however, is that the parameters interact; the effect of any given parameter depends on the other two parameters. The sum of the contributions does not necessarily equal the sum of the overall change, ΔBrS. To solve this problem, parameter changes are computed sequentially. For instance, the contributions due to changing bias first, then information, and finally noise are:

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

These differences form a telescoping sum that equals the overall change, ΔBrS. But the order of the parameters matters. For instance, in the previous 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 would have gotten different contributions, and it is unclear which order should be preferred.

Fortunately, we can use cooperative game theory to solve our problem by treating bias, noise, and information as three team members working together to improve accuracy. In cooperative game theory, the common solution is to average 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 desirable properties (see Hart 1989), including the fact that they sum to the overall change. Appendix C offers an example of how individual contributions are calculated and why order matters.

4. Geopolitical Forecasting Data

4.1. Data

This section applies the BIN model to data from the geopolitical forecasting tournament sponsored by the Intelligence Advanced Research Projects Activity (IARPA) in 2011–2015. It includes hundreds of forecasting questions and probabilistic predictions made by thousands of participants in the Good Judgment Project (GJP). An example question was whether Serbia would be granted European Union (EU) candidacy by December 31, 2011. Forecasting began on September 1, 2011. The question resolved as "no" because Serbia did not gain EU candidacy by December 2011. The question was open for four months. All GJP data are public.[2]

To be included in our analysis, a question had to satisfy two criteria: (i) a binary outcome (yes/no), and (ii) open no more than 180 days. These criteria make questions more comparable. To ensure bias has a consistent interpretation, we standardized the orientation of the outcomes by rescoring 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 underestimation of change.

Forecasters were encouraged to update predictions whenever their beliefs changed. Forecasters knew they would be scored on a Brier metric, and they received a tutorial on the logic behind this scoring rule. If they did not update on a given day, we assumed their beliefs had not changed. We treated their most recent forecasts as their current beliefs about the event. Given that our model produces many results on any given day, it is impractical to analyze all time horizons in depth. Therefore, we present detailed results on predictions made 30 days prior to outcome resolution. This puts all forecasters on the same temporal playing field and ensures at least some uncertainty at the time of their predictions. We repeat the analysis for each of the 60 days before resolution

dates and summarize those 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 research team 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 were advised to consider reference classes; average multiple predictions from different sources; and avoid judgmental biases such as overconfidence, confirmation bias, or base-rate neglect. Manipulation checks verified forecasters had mastered the content. The control group had no training.
- Teaming: Forecasters worked to asynchronous teams in which they could debate each other's predictions. The control group consisted of individuals who worked independently.
- Tracking: Forecasters' performance was 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. These forecasters were not randomly selected; they earned their positions. There is no control group, but we use comparison groups such as regular teams with training.

Our goal is to study how these treatments influence bias, noise, and information by comparing treatment to control (or comparison) groups.

IARPA sought to measure the forecasting equivalent of fluid intelligence, as uncontaminated as possible by specialized knowledge of persons, places, or political processes. To this end, they chose 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. Even though the events were extremely heterogeneous, they all belong to a common reference class, namely changes to the status quo. The BIN model results must then be interpreted in the context of this class of events. The base rate here is the frequency with which change occurs over status quo, and the BIN components (bias, information, noise) shed light on forecasters' skill at predicting change.

Given the heterogeneity in the events, outcomes are unlikely to show much dependence, so the independence part of Assumption 3 seems reasonable. To test the assumption, we calculated the correlation between the predictions of one event and those of another event within each condition of the randomized control trials run by the GJP. 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 correlations were 0.12 and 0.13, respectively. The interquartile range of the correlations was [0.0, 0.25]. This suggests that, on average, Assumption 3 is a fair approximation to the GJP data. Although there were mild positive or negative correlations, 97% of the correlations were within the safe range of our sensitivity analysis.

4.2. Model Validation

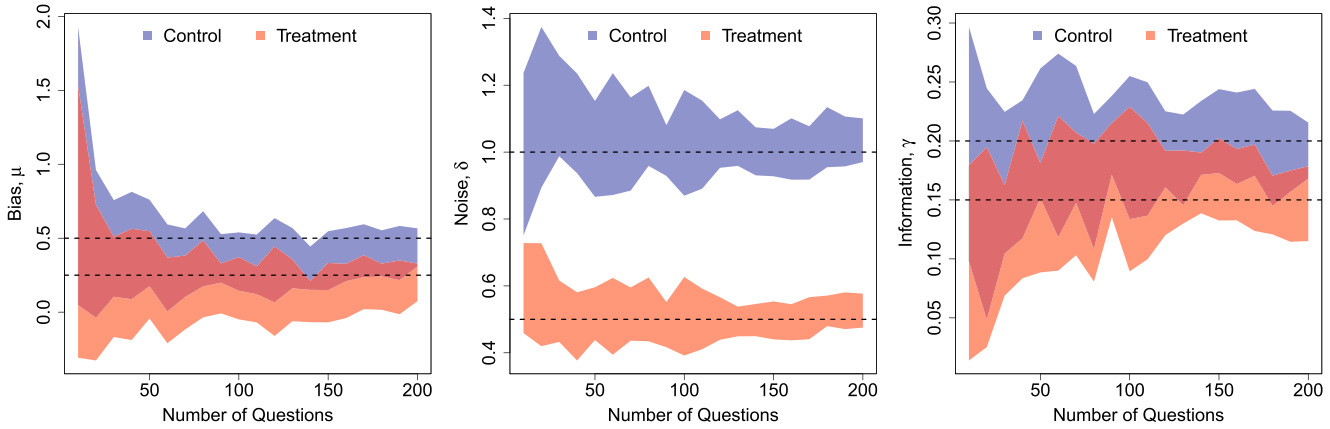
Before we apply the BIN model to the GJP data, we evaluate its potential on simulated data with known parameter values. Ideally, we would perform an exhaustive evaluation over all possible parameter settings. However, our model has 10 parameters that can take on numerous values. Furthermore, the sizes of the groups vary across questions, and the number of questions varies across contexts. An exhaustive evaluation cannot be computed easily or presented succinctly.

We constructed a controlled simulated environment that resembles our application to the GJP data. We based parameter values on estimates of the GJP data. First, we fixed the base rate of change to 0.21. Second, we assumed that each comparison group had 75 forecasters, approximately the median number of forecasters in each condition of the GJP data. Third, we considered a control group that resembled untrained individuals. Specifically, we set $\mu_0 = 0.50$, $\gamma_0 = 0.20$, and $\delta_0 = 1.00$. We then varied the parameters of the treatment group and reported the accuracy of our estimation. Fourth, we set the covariance in the forecasters' interpreted signals to 0.05. Signal interpretations are thus mildly correlated both within and between groups.

We begin with a hypothetical treatment group: $\mu_1 = 0.25$, $\gamma_1 = 0.15$, and $\delta_1 = 0.5$, which reduced bias, noise, and information relative to the control group: $\mu_0 = 0.50$, $\gamma_0 = 0.20$, and $\delta_0 = 1.00$. Figure 3 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 narrow and gravitate toward the true values as the number of questions increases. Figure 3 illustrates the consistency of Bayesian estimation techniques. In the GJP, forecasters in the treatment and control groups answered between 87 and 191 questions. The simulation suggests we have enough data to get reasonable estimates of the parameters.

Although this is only one simulated data set, we can obtain a more sensitive evaluation of our estimation procedure by repeating the analysis on many simulated data sets. In Section 3 of the supplementary

Figure 3. (Color online) Shaded Regions Are the 95% Credible Intervals Under Varying Numbers of Questions



Notes. Blue (higher region) represents the control group and red (lower region) represents the treatment group. Dashed horizontal lines are the true parameter values.

material, we fix the number of questions to 100 and generate 1,000 data sets with varying combinations of $\mu_1 \in \{-0.5, 0\}$, $\gamma_1 \in [0.01, 0.3]$, and $\delta_1 \in [0, 2]$. The results therein show that the root-mean-squared-error in estimating the true parameter values is always below 0.14, and in most cases, well below 0.05, suggesting that our method captures true values accurately across a wide range of parameter values.

In addition to the previous tests, we performed out-of-sample predictive tests on the GJP data. In the absence of a benchmark model, we compared our BIN model against three simplified versions. One removed the noise parameter, one removed the bias, and one assumed information symmetry within each group. We discovered that adding noise or bias parameters improved model fits, whereas allowing informational asymmetry had little effect. Thus, bias and noise are crucial for modeling the GJP data, whereas information asymmetry across groups is less important. Results are presented in Section 9 of the supplementary material.

4.3. Thirty Days to Resolution: Glossary

Tables 1–3 present predictions 30 days prior to outcome resolution and pairwise comparisons of control and treatment groups in each column.[3] Table 1 is divided into the following sections:

- Parameter estimates: We show posterior means of the parameters of interest and their differences. Below each mean is the 95% credible interval that represents the range in which the true parameter value falls with 95% probability. The credible interval differs from the classical 95% confidence interval in that it contains the true parameter value with 95% probability. Because confidence intervals treat parameter values as fixed, the value is either in the interval or not, and hence probabilistic statements are not meaningful. But a Bayesian statistical framework solves that problem.
- Posterior inferences: This section provides the posterior probabilities of events. Compared with 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, 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. In each comparison, both groups made forecasts for the exact same set of questions. 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 years. Furthermore, not every forecaster predicted every question.

Table 1. Parameter Estimates (with 95% Credible Interval) and Posterior Inference for the Good Judgment Project Data

	Training		Teaming		Tracking	
Summary statistics	Individuals: untrained vs. trained	Teams: untrained vs. trained	Untrained: individuals vs. teams	Trained: individuals vs. teams	Trained: teams vs. supers	Untrained individuals vs. supers
Parameter estimates (with 95% CI)						
Outcome mean: μ^*	−0.89 (−1.10, −0.70)	−0.92 (−1.24, −0.62)	−0.94 (−1.24, −0.64)	−0.88 (−1.09, −0.68)	−0.84 (−1.11, −0.59)	−0.84 (−1.09, −0.59)
Bias (control): μ_0	0.46 (0.29, 0.65)	0.42 (0.13, 0.72)	0.51 (0.23, 0.81)	0.42 (0.24, 0.61)	0.22 (−0.00, 0.45)	0.36 (0.14, 0.60)
Bias (treatment): μ_1	0.45 (0.28, 0.63)	0.40 (0.13, 0.69)	0.44 (0.17, 0.73)	0.31 (0.14, 0.48)	−0.09 (−0.29, 0.13)	−0.09 (−0.30, 0.12)
Difference in bias: $\lvert\mu_0\rvert - \lvert\mu_1\rvert$	0.01 (−0.01, 0.04)	0.02 (−0.03, 0.06)	0.07 (0.01, 0.13)	0.11 (0.07, 0.16)	0.11 (−0.26, 0.35)	0.25 (−0.15, 0.51)
Information (control): γ_0	0.34 (0.27, 0.40)	0.29 (0.15, 0.41)	0.21 (0.08, 0.33)	0.31 (0.23, 0.38)	0.44 (0.34, 0.52)	0.32 (0.22, 0.41)
Information (treatment): γ_1	0.34 (0.27, 0.40)	0.35 (0.23, 0.44)	0.32 (0.19, 0.42)	0.45 (0.38, 0.51)	0.60 (0.52, 0.66)	0.60 (0.52, 0.67)
Difference in information: $\gamma_0 - \gamma_1$	−0.00 (−0.03, 0.03)	−0.05 (−0.11, 0.00)	−0.10 (−0.17, −0.04)	−0.14 (−0.18, −0.10)	−0.15 (−0.22, −0.10)	−0.27 (−0.35, −0.20)
Noise (control): δ_0	0.91 (0.76, 1.09)	0.80 (0.54, 1.16)	1.07 (0.78, 1.43)	0.77 (0.62, 0.96)	0.65 (0.46, 0.91)	1.02 (0.78, 1.32)
Noise (treatment): δ_1	0.70 (0.57, 0.86)	0.51 (0.33, 0.77)	0.74 (0.51, 1.06)	0.54 (0.42, 0.70)	0.28 (0.15, 0.47)	0.28 (0.15, 0.48)
Difference in noise: $\delta_0 - \delta_1$	0.21 (0.13, 0.29)	0.28 (0.14, 0.45)	0.33 (0.16, 0.50)	0.23 (0.13, 0.33)	0.37 (0.21, 0.55)	0.74 (0.54, 0.95)
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	0.73	0.90
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: $\mathbb{P}(\gamma_0 < \gamma_1)$	0.52	0.97	1.00	1.00	1.00	1.00

Notes. 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 nontrained individuals and those who are already trained and working in teams. CI: credible interval.

Table 2. Analysis of Predictive Performance for the Good Judgment Project Data

	Training		Teaming		Tracking	
Summary statistics	Individuals: untrained vs. trained	Teams: untrained vs. trained	Untrained: individuals vs. teams	Trained: individuals vs. teams	Trained: teams vs. supers	Untrained individuals 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%	3.7%	6.0%	10.0%	15.0%
Increase in information	0.0%	1.3%	2.0%	3.8%	6.4%	8.1%
Reduction in noise	6.2%	10.0%	8.3%	8.1%	16.9%	23.5%
Maximum achievable contribution						
Reduction in bias	33.0%	30.4%	34.5%	31.0%	16.5%	24.5%
Increase in information	16.0%	19.2%	15.3%	19.6%	26.7%	20.2%
Reduction in noise	50.3%	49.8%	49.6%	48.7%	55.9%	54.7%

4.4. Thirty Days to Resolution: Discussion

Posterior estimates in Table 1 show that the base rate of change is $\Phi(\mu^*) \approx \Phi(-0.88) = 0.189$ and that all groups, except the trained and superteam conditions, had upward bias or assigned excessive probabilities to departures from the status quo. Second, the amount of information varied across groups (from 0.21 to 0.60, with an information level of 1.0 being equivalent to omniscience). Third, all groups had noise, and some had much more than others.

Super teams were the most informed, least noisy, and least biased. Superforecasters had been selected based on their prior excellent performance, so their information, bias, and noise can been seen as rough approximations of what is achievable—by humans—within this forecasting environment.

Posterior inferences show how much treatment groups outperformed control groups on each component. First, all treatments reduced noise with an estimated posterior probability of virtually 1.0.[4] Second, all treatments, except probability training of individuals, increased information. Posterior probabilities ranged from 0.97 to 1.00. Third, only teaming reduced bias. Posterior probabilities ranged from 0.99 to 1.00. Although our methodology can be used to compare the magnitude of positive or negative bias separately across groups, we focus on the absolute value of the bias for simplicity: $|\mu_0|$ and $|\mu_1|$. So, a forecaster with, for example, $\mu_0 = -0.1$ is more biased than one with $\mu_1 = 0.05$.

Our results suggest that reducing noise is easier than reducing bias. Noise was tamped down via training, teaming, or tracking, whereas bias was only reduced by teaming. This finding is not as surprising as it may initially sound because teams were warned about groupthink and told that, to maximize accuracy, they should be actively open-minded and grapple with dissonant arguments.

We suspect that training improved forecasters' skill by giving them a more granular understanding of uncertainty, which helped them translate vague verbal hunches like "distinct possibility" and "very likely" into quantitative values (Friedman et al. 2018). Insofar as trained individuals were better at recognizing that

Table 3. Summary Statistics for the Good Judgment Project Data

	Training		Teaming		Tracking	
Data summary	Individuals: untrained vs. trained	Teams: untrained vs. trained	Untrained: individuals vs. teams	Trained: individuals vs. teams	Trained: teams vs. supers	Untrained individuals vs. supers
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

forecast of 0.95 implies 19:1 betting odds and 0.9 implies much lower betting odds of 9:1, we should expect their judgments to be more consistent, and less noisy.

Not surprisingly, forecasters who invest more cognitive effort in their predictions report probabilities that better reflect their true beliefs (information). But there are different ways to encourage cognitive effort and different paths by which greater effort could translate into accuracy. Teaming is one motivator. Forecasters working in teams can see each other's predictions, which introduces accountability. Extreme predictions in the wrong direction can cause a forecaster to lose status. As a result, team members tend to make more circumspect and less erratic predictions, 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 surprising to find less noise in their judgments.

Reducing bias, however, may be harder than reducing noise. Even though bias can be caused by many factors, one plausible explanation is that adding zero mean noise to the forecasters' interpretations makes predictions too extreme and overconfident in the sense that they are often too close to the endpoints of zero and one (Erev et al. 1994). Forecasters can reduce noise by being more conservative and avoiding extreme predictions. Reducing bias, however, is more challenging because it may require forecasters to estimate the base rate. To do this, they need to define a reference class and collect information on these events, both of which require skill and effort. Forecasters grappled with base rates as diverse as the number of African countries with dictators, time series of Nikkei stock prices, frequencies of United Nations Security Council arms embargos, and variations in euro-dollar exchange rates.

Our results, however, suggest that cognitive biases here are not as irresistible as perceptual illusions (Arkes 1991, Lerner and Tetlock 1999). First, superforecasters showed relatively little bias. Second, teaming reduced bias, which may reflect how teams were instructed to interact—by second-guessing each other to avoid excessive conformity/herding (Tetlock and Gardner 2016). Consistent with past work, properly organized and incentivized groups can check at least certain cognitive biases (e.g., Kerr et al. 1996).

Posterior probabilities suggest that the interventions decreased bias and noise and increased information, but statistical and substantive significance are not the same thing. To understand the substantive importance of each change, the percentage of control group Brier scores in Table 2 provides the normalized individual contributions, as in Section 4.3. These data reaffirm that noise reduction is most important to improving accuracy. Training improves accuracy almost entirely through noise reduction. 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 organizations that care about forecasting accuracy. Inaccuracies are 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, by roughly 25%; and increasing information would deliver the remaining 25%. In sum, reducing noise is roughly twice as effective as reducing bias or increasing information.

Given that our analysis is based on a statistical model, there is always the risk that the noise term is capturing unmodeled systematic effects or model misfits. The BIN model might better fit the treatment than the control group, creating the misimpression that the treatment reduces noise. Although we cannot rule out such possibilities completely, we can render them implausible by testing across a variety of treatments and by testing the model on a variety of simulated and experimental datasets (see Sections 3 and 5 in the supplementary material).

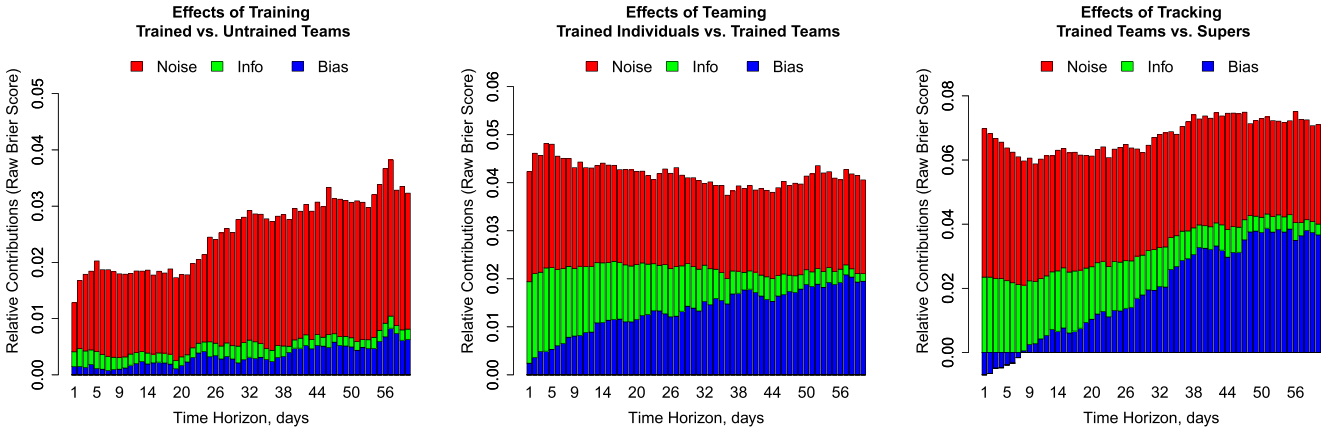
4.5. Other Horizons

Space constraints prevent us from showing Tables 1–3 for all time horizons. But Figure 4 can show the waxing and waning of the relative contributions of bias, noise, and information to the differential Brier scores of the control and treatment groups from days one to 60 (with more data in Section 6 of the supplementary material). For any given day, bars represent differences in expected Brier scores due to noise, information, and bias. The sum of the 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 shows that noise reduction is the main reason that trained teams outperform untrained teams. Noise effects are bigger than bias or information. Noise effects are also larger at longer (30 to 60 days) than shorter horizons (one to 30 days). Bias reduction contributes more to performance at longer horizons. The contribution of information does not change much over time.

The center panel shows the effects of teaming on trained forecasters. Again, noise reduction is the main reason that teams outperform individuals. Noise reduction drives almost half of the difference in Brier scores, an effect that remains constant over time. Bias and information share the remaining portions and move in opposing directions. Teams reduce bias more at longer horizons and boost information at shorter

Figure 4. (Color online) Contributions of Noise, Information, and Bias in Reducing Brier Scores at Varying Time Horizons



horizons. As time unfolds, more information becomes available, and teaming—unlike training in the left panel—allows forecasters to harness the information.

The right panel presents the contributions of tracking elite teams (superforecasters) versus trained regular teams. Relative contributions resemble the pattern for teaming in the center panel. At day 60, elite teams have less noise and less bias than regulars. Information differences are not large. As the resolution date draws near, elite teams have less noise, more information, and slightly more bias than regulars. And there is a negative contribution of bias for superforecasters around 10 days before the resolution date, the only occasion when super teams are more biased than regular teams. Close inspection reveals that both averages begin above the base rate but average forecasts among superforecasters fall below the base rate, whereas average forecasts among regulars do not. As the horizon shrinks (less than 10 days), superforecasters become more biased than regulars. Why? We suspect that top performers may have tried too hard to win by assigning forecasts of 0.0 even though a week still remained for surprises. This is a response bias (not a perceptual bias) that is presumably driven by the superforecaster-status incentives in the tournament.

5. Conclusion

We divide our closing remarks into three categories: (a) What have we learned from applying the BIN model to experiments embedded within forecasting tournaments? (b) Does the BIN model facilitate discoveries that would have been harder using alternative models and techniques? (c) What should be the priorities for further empirical work?

5.1. Key Discoveries and Their Implications

Of the three methods of boosting accuracy—tamping down noise and bias or ramping up signal detection—noise reduction emerged as the most consistent driver across the three experimental interventions: training forecasters in probabilistic reasoning, organizing them into open-minded teams, and tracking the best forecasters into elite teams. Interventions that boosted accuracy also suppress random errors in judgment, even when the original intent of the investigators was to do so via other mechanisms. For instance, the training was originally envisioned as a debiasing manipulation that would reduce biases, such as base-rate neglect, by encouraging forecasters to adopt the outside view (Kahneman 2011). The rationale for teaming was originally to prevent biases such as groupthink and failures to pool information. And the original rationale for tracking was to assemble the most insightful signal detectors into teams and see whether these elite teams would be more accurate than regular teams (Mellers et al. 2015).

These findings raise new questions. If noise reduction is indeed the key driver across disparate experimental interventions, what drives noise reduction? Can we isolate which facets of multifaceted interventions tamped down noise? In retrospect, it makes sense that debiasing training, which stresses the value of grounding initial probability judgments in base rates, would also have the net effect of stabilizing forecasts. It makes sense that teaming would help forecasters converge on more reliable crowd judgments (e.g., Farrell 2011). And it makes sense that tracking top forecasters into elite teams would facilitate convergence on even more reliable crowd aggregates. But we should be wary of hindsight bias. For these themes did not loom large in prior publications from the GJP, which emphasized the capacity of training to tamp down bias (Chang et al. 2016) and the skills of superforecasters as subtle signal detectors (Mellers et al. 2015).

We focus here on the practical and theoretical significance of two discoveries the BIN model has revealed about top performers in forecasting tournaments.

First, the results place qualifications on past portrayals of top performers (superforecasters) in both the scientific literature (Mellers et al. 2015) and in popular books (Tetlock and Gardner 2016). Earlier work had stressed the insightful ways in which top performers either extracted subtle predictive clues that others missed or avoided being gulled by pseudodiagnostic cues that others were misled into using. Our data suggest that superforecasters owe their success more to superior skills at tamping down measurement error, than to unusually incisive readings of the news. Discipline may matter more than creativity here.

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, which tempts them to make overconfident claims. The decision calculus for top performers may take this form: "If I adjust my forecast from .05 to 0.005 even though I believe the probability is 0.05, I will have a slightly better Brier score (0.000025 vs 0.0025) but I will have that benefit on the vast majority of 0.05 probability events that do not materialize (99% of the events for which the supers' average forecast five days before resolution were less than 0.05, did not occur). Those tiny advantages will accumulate and put me in first place. Of course, there would be a steep scoring penalty if these almost-slamdunk events actually materialize (the remaining 1% of events). That would degrade my accuracy score. I might even fall into the lower ranks. But the risk is worth taking for a shot at being 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. Even though competition among forecasters can lead to less accurate probabilities, there are exceptions (Pfeifer 2016). We cannot say for sure that this decision calculus drives the bias bump that the BIN model discovered. But it is a parsimonious explanation for the bump among superforecasters toward the end of forecasting periods.

5.2. Alternative Analytic Models and Decompositions: Could We Have Learned the Same Things from Them?

Our framework connects to other cognitive models, most notably the Brunswikian lens model (e.g., Karelaia and Hogarth 2008), in which both outcomes and forecasters' predictions depend linearly on a finite set of predictive cues. The "achievement index" and "consistency" with which a forecaster executes the decision rule relate to information and noise in our BIN model. One common use of the lens model studies whether forecasters' subjective cue weights correspond to the objective or true weightings of those cues in the environment. To calculate these weights, however, researchers must know a lot more about the forecasters and the world than is often possible (or was possible in the geopolitical tournaments). For this reason, it is much easier to apply the lens model in laboratory settings than in tournaments. The BIN model allows each forecaster to integrate a large (in principle, infinite) number of cues into a probability judgment in ways that need not be known to the researcher. All that matters is the final judgment, not how the forecaster reached it (for more details, see Section 1 in the supplementary material).

The BIN model also captures key aspects of forecasting in Bayesian modeling of subjective probability judgments (Clemen and Winkler 1999), including precision, bias, and dependence. Precision is forecasters' skill at discriminating occurrences from nonoccurrences, which corresponds in the BIN model to the information parameter γ. The more informed the forecaster, the more precisely the can predict the event.

Dependence is forecasters' tendency to make similar predictions for an event, which corresponds in the BIN model to the covariance of forecasters' interpretations. When predictions target the same outcome, correlations can stem from either shared information or shared misconceptions. The BIN model treats the magnitude of the within-event correlations as an empirical matter. Furthermore, we model the within-event dependence separately for each group and across groups. For example, the dependence within the control group need not equal the dependence in the treatment group or the dependence between forecasters across groups. These parameters are not of direct interest here but we discuss them in Appendix A of the main text and Section 7 of the supplementary material.

Clemen and Winkler (1999) define bias as any type of miscalibration that is typically inspected by calibration plots that display predictions against empirical frequencies of events. For a recent discussion on how to interpret and estimate calibration plots from data, see Attali et al. (2020). If a forecaster is perfectly calibrated, all points fall on the diagonal—and deviations imply miscalibration. The BIN model defines bias as systematic over- or underestimation of probabilities and changes the vertical level of the points in a calibration plot. However, miscalibration can also arise from under- or overconfidence: the forecaster's predictions are too close too, or far from, the base rate. In practice, forecasters are often overconfident (e.g., De Bondt and Thaler 1995).

Even though noise is typically defined as random error, it can have systematic effects on the predictions. As Erev et al. (1994) explain, systematic overconfidence can be introduced by adding zero mean noise to calibrated predictions, which inflates the variance of the prediction and makes it deviate too much from the base rate. Similarly, the BIN model can capture overconfidence: by increasing the noise level δ, and making predictions overconfident, which changes the slope of the points in a calibration plot. Via these mechanisms, the BIN model can capture many forms of miscalibration. However, following Gigerenzer (2018), the model treats under- and overconfidence not as a bias but as noise, and reserves bias to refer to systematic over- or underestimation of probabilities.

The most important aspect of the BIN model, however, is the final decomposition of the Brier score. Our decomposition differs from previous decompositions in several ways (e.g., Murphy 1973, Yates 1982, Murphy and Winkler 1987, Davis-Stober et al. 2015). First, previous work (e.g., Bröcker 2009) has focused on decomposing proper scoring rules into calibration (how well probabilities align with empirical frequencies) and resolution (the extremity of the probabilities). The BIN model can decompose any rule, including proper scoring rules such as logarithmic and spherical scoring (e.g., Gneiting and Raftery 2007) as well as improper scoring rules such as the absolute deviation. Regardless of the chosen metric, the BIN decomposition always brings us back to bias, information, and noise, whose interpretations do not change and are easy to understand. Second, previous decompositions have been done on the probability scale. The unavoidable price of ensuring that probability predictions stay inside the unit interval (and never fall below zero or rise above one) is that estimates of bias and noise become logically intertwined: the larger the bias, the smaller the noise must be. This confounding makes it difficult, arguably impossible, on a bounded probability scale, to separate bias from noise. For this reason, the BIN model estimates components on an unbounded probit scale. Third, existing decompositions do not generate uncertainty estimates of the components or permit statistical comparisons of treatments. Although uncertainty could be estimated by bootstrapping techniques, the validity of such an approach would depend on technical conditions that are nontrivial to verify (e.g., Wasserman 2006, theorem 3.20). By contrast, our approach permits valid comparisons based on Bayesian statistics and a joint probabilistic model of all outcomes and predictions.

5.3. Limitations and New Research Directions

In principle, the BIN model could be extended to estimate bias, information, and noise separately for each individual. This could be achieved by modeling components in terms of covariates, such as the forecasters' gender, age, self-reported expertise, or the forecast horizon, or via a hierarchical approach that models each forecaster's bias, noise, and information as draws from a common distribution. In practice, however, modeling bias, information, and noise at an individual level requires a dense covariance matrix of forecasters' interpretations that captures the allocation of information among forecasters. These allocations must be logically feasible. For instance, two forecasters cannot each know more than 50% of all information but not share any information. Ensuring an allocation is jointly feasible for a group of forecasters is combinatorially challenging. Although the partial information constraints can be managed under the symmetric model in this article, in the general case the covariance matrix must belong to the so-called correlation polytope (Satopää et al. 2016, proposition 3.1) or at least its semidefinite relaxation (Satopää et al. 2017). Unfortunately, it is not clear at this point how we can estimate a covariance matrix within these constraints. Even though we believe that such models are attainable, they will require significant technological innovation.

Future work could also relax the (conditional) independence assumption. Although our estimation procedure remains robust under mild to moderate levels of dependence, in some contexts, such as macroeconomic forecasts, the events may be highly dependent manifestations of a common underlying process. Here there is unlikely to be a one-size-fits-all solution. For instance, dependence can be spatial (e.g., hurricane trajectories) or temporal (e.g., macroeconomic indicators). Only by adapting the model to the application at hand can one make sound inferences and arrive at reliable results.

Incorporating dependencies is an exciting goal for future work but it raises serious practical challenges with human forecasters. Asking people to predict dependent events reduces the effective sample size of questions (and increases workload) and makes it statistically harder to separate true from spurious effects. As an analogy, consider the testing of students. To evaluate their overall understanding of the course, we would use independent questions, not questions where the correct answer to one question hints at the correct answer to another. GJP questions were by design different and hence are less correlated, making it appropriate for testing the forecasters' general forecasting skill under different treatments.

Overall, noise reduction emerged as a key driver of the Good Judgment interventions to improve forecasting, even when the designers of the interventions did not have noise reduction in mind. So it is natural to ask how much more important noise reduction

might become when it is the research objective. We see at least four lines of work that could be usefully linked to the BIN model and related frameworks:

a. Disciplining the internal judgment processes of forecasters by requiring them to participate in noise audits (Kahneman et al. 2016) or other training exercises (Chang et al. 2016) that gauge how similarly or differently they perform highly structured tasks in which there is no ambiguity about which cues are informative and which incentives/instructions are designed to be sources of bias.

b. Aggregating judgments either through institutional interventions such as prediction markets (Wolfers and Zitzewitz 2004, Atanasov et al. 2017) or statistical means (Larrick and Soll 2006, Budescu and Chen 2014, Satopää et al. 2014, Prelec et al. 2017). For instance, does crowd averaging operate mainly as a noise reducer (as classically supposed) but more complex algorithms deliver benefits via more complex pathways? One of our ongoing projects analyzes the effects of a variety of forecast aggregators, including averaging and prediction markets, on bias, information, and noise. One of the top-performing aggregators is based on the BIN model. This aggregator gains its superior performance almost entirely through improved calibration, suggesting that its underlying model, namely the BIN model, is a more appropriate description of the forecast-generating process than the models underlying other commonly used aggregators.

c. Interventions aimed at simplifying the external world by filtering out misleading or low-diagnosticity sources in the news environment and lightening the cognitive load on forecasters (Lazer et al. 2018).

d. The most radical of measures, replacing human judges with machine-learning algorithms, as has been done already in many domains, such as targeting restaurants for health inspections (Kang et al. 2013) and identifying teenagers at highest risk for committing crimes (Chandler et al. 2011).

Of course, it is possible to have too much of a good thing. We should factor in the potential costs of these targeted noise-reduction interventions, including the risks of inducing too much uniformity in how forecasters interpret signals. That said, noise may still prove to be the easiest source of error to correct, and organizations looking for cost-effective methods of boosting accuracy should give noise-reduction tools serious consideration.

Finally, we acknowledge the need for work on the BIN model to iterate between real-world settings emphasized here and controlled laboratory settings in which researchers can manipulate the diagnosticity of the informative cues, forecasters' temptations to make biased forecasts, and the stochasticity of the environment. We report extensive statistical simulation data in the supplementary material and also report an illustrative human subject demonstration of the advantages of experimentally manipulating BIN model components (see Sections 3 and 5 in the supplementary material, respectively). In particular, using an iterated prisoner's dilemma (iPD) game played by bots programmed to implement strategies varying from competitive to cooperative, we challenged human forecasters to predict patterns of play under four sets of conditions:

a. Control condition in which forecasters had relatively little information, and faced relatively low levels of systematic (bias) and random (noise) distortions in the shown rounds of play.

b. Information condition in which forecasters saw roughly twice as many rounds of play, giving them more opportunities to detect patterns in bot strategies.

c. Noise condition in which forecasters have the same information set as the control group but had to cope with the confusion induced by randomly changing three defections to cooperation or three cooperative acts to defection.

d. Bias condition in which forecasters had the same information as the control group but the misperception errors were introduced asymmetrically, causing cooperation to become defection but not the reverse.

We present the data in more detail in the supplementary material but, suffice it to say here, the demonstration pilot study worked: the BIN model-estimated posterior probabilities captured the experimental manipulations of the forecasting environment.

Field experiments (such as forecasting tournaments) and laboratory experiments (of the iPD sort) are deeply complementary methods of testing the BIN model and fine-tuning methods of improving forecasting. Laboratory simulation studies actually let us rerun history and assess the true probability distributions of possible worlds at each turn in the game, and neutralize the standard objection to putting probability estimates on the ostensibly unique or one-off events in IARPA tournaments (Tetlock 2017). The strongest tests of the BIN model would show that the same methods improve forecasting work in both simulated worlds and in the real world, and do so via similar underlying mechanisms.

That said, we should not understate what we can learn from real-world forecasting tournaments. Let's assume, for the sake of argument, that 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 either 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 (as in simulations) and measure how closely forecasters' estimated distributions of possible worlds correspond to actual distributions. But the singular-event objection does not apply even to the real-world tournament data. For we have applied the BIN model to predictions of multiple singular events, or to be precise, to a set of ostensibly singular events within which it proved 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 never have been possible to identify systematic individual difference correlates of accuracy or experimental interventions that boost accuracy. The BIN model shows that the standard objection over-reaches and needs reformulation. More generally, the BIN model provides a method for understanding the mechanisms behind interventions intended to improve accuracy in both artificial and real worlds. Knowing how these interventions work makes it easier to see what still needs to be done for even greater accuracy.

Acknowledgments

The U.S. Government is authorized to distribute reprints notwithstanding any copyright annotation. The views expressed here are those of the authors, not of IARPA or the U.S. Government.

Appendix 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

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

where

$$\begin{aligned} \boldsymbol{\Sigma}_{gg,k} &= \mathrm{Cov}(\mathbf{Z}_{g,k}, \mathbf{Z}_{g,k}) \\ &= \mathbf{I}_{N_{g,k}}(\delta_g + \gamma_g - \rho_g) + \mathbf{J}_{N_{g,k} \times N_{g,k}} \rho_g \text{ for } g \in \{0,1\}, \\ \boldsymbol{\Sigma}_{g,k} &= \mathrm{Cov}(\mathbf{Z}_{g,k}, Z_k^*) = \mathbf{1}_{N_{g,k}} \gamma_g \text{ for } g \in \{0,1\}, \\ \boldsymbol{\Sigma}_{01,k} &= \mathrm{Cov}(\mathbf{Z}_{0,k}, \mathbf{Z}_{1,k}) = \mathbf{J}_{N_{0,k} \times N_{1,k}} \rho_{01}, \\ Y_k &= \mathbf{1}(Z_k^* > 0), \\ p_{g,j} &= \Phi\left(\frac{Z_{g,j}}{\sqrt{1-\gamma_g}} \right) \text{ for } j = 1, \ldots, N_{g,k} \text{ and } g \in \{0,1\}, \end{aligned}$$

$\mathbf{J}_{a \times a}$ 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

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

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, vague prior distribution on $\boldsymbol{\theta}$; that is, $\pi(\boldsymbol{\theta}) \propto 1$. The parameters are then estimated with a MCMC technique called Hamiltonian Monte Carlo sampling. This is a standard estimation procedure in Bayesian statistics that we implement using the Stan software package (Carpenter et al. 2017). In each case, the algorithm ran 4,000 iterations, of which the first 2,000 were used for burn-in. See Section 10 in the supplementary material for further convergence diagnostics under the GJP data. 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.

Our implementation is available in the supplementary material and allows the user to easily modify the prior distributions. This can be useful if the user has prior knowledge about the plausible ranges of the parameters. To provide guidance, consider the information parameter γ_g in group $g \in \{0,1\}$. This is always between 0 (no information) and 1 (full information). In practice, it is unlikely that γ_g is very close to 1, especially if the forecast horizon is high. The user may then consider a beta prior distribution that places relatively more weight on smaller values of γ_g. Similarly, even though the noise parameter, δ_g, is unbounded from above, it is unlikely that δ_g is much higher than 1. To provide intuition, $\delta_g = 1$ implies that the forecasters have, in expectation, as many irrelevant signals as there are relevant signals in the signal universe. If so, forecasters have a very high and unlikely level of noise. The user may then consider an exponential prior distribution that places relatively more weight on smaller values of δ_g. A similar argument can be made about the bias parameter μ_g. If the base rate is not very close to 0 or 1, then μ^* is close to 0. As a result, a bias in $[-2, 2]$ or, say, in $[-5, 5]$ can capture most differences between the base rate and the forecasters' systematic bias. The user may then consider a Gaussian prior distribution that places relatively more weight on values of μ_g near 0.

The exact choice of prior is less important if we use a reasonably large data set. With more data, the influence of the prior on the posterior is likely to be "washed away" as long as the model is well-specified. To illustrate, we replicated the analysis of the GJP data with informative priors and the results remained essentially identical to the ones presented in Section 4.3. See Section 8 in the supplementary material for the results.

Appendix 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* $\boldsymbol{\mu}' = [\mu^* \;\; \frac{\mu^*+\mu_g}{\sqrt{1-\gamma_g}}]$ *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:

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

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 $\mathbb{E}_{Z^*,Z_g,\varepsilon}\{\mathbb{P}(\frac{Z_g}{\sqrt{1-\gamma_g}} - \varepsilon > 0)\mathbb{P}(Z^* > 0)\}$. This value is equal to the probability that a bivariate normal random variable given by $[\,Z^* Z_g - \varepsilon\,]$ 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 independent standard Gaussians ε_1 and ε_2. The mean and covariance matrix of the resulting random variable are $\boldsymbol{\mu}''$ and $\boldsymbol{\Omega}''$, respectively. □

Appendix C. Example with Shapley Values

The following stylized example illustrates how the Shapley value is used in calculating the individual contributions.

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

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

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

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 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 $\delta_0 \rightarrow \delta_1 = 0$ decreases the Brier score from 0.5 to 0.25 and, conditional on $\delta_0 = \delta_1 = 0$, changing $\gamma_0 \rightarrow \gamma_1 = 1$ 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.

Endnotes

[1] In our model, the forecaster classifies signals as either noise or information, suppresses all signals believed to be noise, and uses the remaining signals to make a final prediction. A promising future direction is to explore the forecaster's subjective uncertainty in classifying signals as noise or information.

[2] The data can be downloaded at https://dataverse.harvard.edu/dataverse/gjp.

[3] For convergence diagnostics of our MCMC estimation procedure, see Section 10 in the supplementary material.

[4] 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.

References

Arkes HR (1991) Costs and benefits of judgment errors: Implications for debiasing. *Psych. Bull.* 110(3):486–498.

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

Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, Ungar L, Mellers B (2017) Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Sci.* 63(3):691–706.

Attali Y, Budescu D, Arieli-Attali M (2020) An item response approach to calibration of confidence judgments. *Decision* 7(1):1–19.

Bliss CI (1934) The method of probits. *Science* 79(2037):38–39.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Brighton H, Gigerenzer G (2015) The bias. *J. Bus. Res.* 68(8):1772–1784.

Bröcker J (2009) Reliability, sufficiency, and the decomposition of proper scores. *Quart. J. Roy. Meteorological Soc.* 135(643):1512–1519.

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

Budescu DV, Por HH, Broomell SB, Smithson M (2014) The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change* 4(6):508–512.

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. *j. statist. software* 76(1):1–32, doi:10.18637/jss.v076.i01.

Chandler D, Levitt SD, List JA (2011) Predicting and preventing shootings among at-risk youth. *Amer. Econom. Rev.* 101(3):288–292.

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 tournaments. *Judgment Decision Making* 11(5):509–526.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Davis-Stober CP, Budescu DV, Broomell SB, Dana J (2015) The composition of optimally wise crowds. *Decision Anal.* 12(3):130–143.

De Bondt WF, Thaler RH (1995) Financial decision-making in markets and firms: A behavioral perspective. Jarrow RA, Maksimovic V, Ziemba WT, eds. *Handbooks in Operations Research and Management Science*, vol. 9 (Elsevier, Amsterdam) 385–410.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over-and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–527.

Everett B (2013) *An Introduction to Latent Variable Models* (Springer Science & Business Media).

Farrell S (2011) Social influence benefits the wisdom of individuals in the crowd. *Proc. Natl. Acad. Sci. USA* 108(36):E625.

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. *Internat. Stud. Quart.* 62(2):410–422.

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

Gigerenzer G (2018) The bias bias in behavioral economics. *Rev. Behavioral Econom.* 5(3–4):303–336.

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

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102(477):359–378.

Hart S (1989) Shapley value. Eatwell J, Milgate M, Newman P, eds. *The New Palgrave: Game Theory* (The Macmillian Press Limited, London), 210–216.

Henning G (1989) Meanings and implications of the principle of local independence. *Language Testing* 6(1):95–108.

Kahana MJ, Aggarwal EV, Phan TD (2018) The variability puzzle in human memory. *J. Experiment. Psych. Learn. Memory Cognition* 44(12):1857–1863.

Kahneman D (2011) *Thinking, Fast and Slow* (Farrar, Straus & Giroux, New York).

Kahneman D, Rosenfield AM, Gandhi L, Blaser T (2016) Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Bus. Rev.* 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. Yarowsky D, Baldwin T, Korhonen A, Livescu K, Bethard S, eds. *Proc. 2013 Conf. Empirical Methods Natl. Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), 1443–1448.

Karelaia N, Hogarth RM (2008) Determinants of linear judgment: A meta-analysis of lens model studies. *Psych. Bull.* 134(3):404–426.

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 Quart.* 82(2):279–299.

Kerr NL, MacCoun RJ, Kramer GP (1996) Bias in judgment: Comparing individuals and groups. *Psych. Rev.* 103(4):687–719.

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

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

Lee MD (2018) Bayesian methods in cognitive modeling Wixted J, Wagenmakers E-J, eds. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (John Wiley & Sons, Hoboken, NJ) vol. 5, 37–84.

Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psych. Bull.* 125(2):255–275.

McCullagh P, Nelder JA (1989) *Generalized Linear Models*, 2nd ed. (Chapman & Hall, London).

Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, et al. (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspect. Psych. Sci.* 10(3):267–281.

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psych. Sci.* 25(5):1106–1115.

Murphy AH (1973) A new vector partition of the probability score. *J. Appl. Meteorology* 12(4):595–600.

Murphy AH, Winkler RL (1987) A general framework for forecast verification. *Monthly Weather Rev.* 115(7):1330–1338.

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

Pfeifer PE (2016) The promise of pick-the-winners contests for producing crowd probability forecasts. *Theory and Decision* 81(2): 255–278.

Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532–535.

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

Satopää VA, Pemantle R, Ungar LH (2016) Modeling probability forecasts via information diversity. *J. Amer. Statist. Assoc.* 111(516): 1623–1633.

Satopää VA, Jensen ST, Pemantle R, Ungar LH (2017) Partial information framework: Model-based aggregation of estimates from diverse information sources. *Electronic J. Statist.* 11(2):3781–3814.

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

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 University Press, Princeton, NJ).

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

Van Der Bles AM, van der Linden S, Freeman AL, Spiegelhalter DJ (2020) The effects of communicating uncertainty on public trust in facts and numbers. *Proc. Natl. Acad. Sci. USA* 117(14): 7672–7683.

Wasserman L (2006) *All of Nonparametric Statistics* (Springer Science & Business Media, New York).

Wolfers J, Zitzewitz E (2004) Prediction markets. *J. Econom. Perspect.* 18(2):107–126.

Wyart V, De Gardelle V, Scholl J, Summerfield C (2012) Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron* 76(4):847–858.

Yates JF (1982) External correspondence: Decompositions of the mean probability score. *Organ. Behav. Human Performance* 30(1):132–156.