Robust recalibration of aggregate probability forecasts using meta-beliefs

Cem Peker[1] and Tom Wilkening[2]

[1]Center for Behavioral Institutional Design, New York University Abu Dhabi

[2]Department of Economics, Universiy of Melbourne

July, 2023

Abstract

Previous work suggests that aggregate probabilistic forecasts on a binary event are often conservative. Extremizing transformations that adjust the aggregate forecast away from the uninformed prior of 0.5 can improve calibration in many settings. However, such transformations may be problematic in decision problems where forecasters share a biased prior. In these problems, extremizing transformations can introduce further miscalibration. We develop a two-step algorithm where we first estimate the prior using each forecasters' belief about the average forecast of others. We then transform away from this estimated prior in each forecasting problem. Evidence from experimental prediction tasks suggest that the resulting average probability forecast is robust to biased priors and improves calibration.

1 Introduction

Problems of practical decision making typically require probabilistic forecasts on uncertain scenarios and events. Individual forecasters are often miscalibrated due to various cognitive biases or errors (Kahneman et al., 1982; Erev et al., 1994). Combining independent judgments from many forecasters can lead many individual-specific errors to cancel out leading to improved forecasts via the "wisdom of crowds" effect (Larrick & Soll, 2006; Surowiecki, 2004). However, it does not necessary resolve all issues. In particular, aggregated forecasts tend to be too conservative with the probability of unlikely events being over-predicted and the probability of near-certain events being under-predicted (Ariely et al., 2000; Turner et al., 2014).

There are a variety of explanations for why the aggregated probability forecasts are conservative. First, there is strong evidence that forecasters overestimate the chance of rare events and underestimate highly likely events (Camerer & Ho, 1994; Fischhoff et al., 1977; Moore & Healy, 2008; Wu & Gonzalez, 1996). Since such biases are systematic, they are unlikely to disappear in the aggregate. Second, there are potential issues of censoring on the boundary of the probability space if judgment errors are symmetric and additive. Such censoring issues will naturally lead to a bias towards 0.5 (Erev et al., 1994; Baron et al., 2014). Finally, forecasters may anticipate that they have access to relatively small amounts of information compared to the total information available. If they are not confident in their information, they may naturally report predictions that are too close to their prior (Baron et al., 2014).

One way to address the conservative bias and improve calibration is to extremize the aggregate probability. Consider the linear log odds (LLO) transformation

$$t(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}, \tag{1}$$

where p and $t(p)$ are the original and transformed probabilities, and $\{\delta, a\}$ are parameters

(Turner et al., 2014). The LLO transformation follows from a linear log-odds model

$$log\left(\frac{t(p)}{1-t(p)}\right) = \gamma log\left(\frac{p}{1-p}\right) + \tau, \qquad (2)$$

where γ is the slope and $\tau = log(\delta)$ gives the intercept (Turner et al., 2014)[1]. Extremizing transformations of the LLO form typically improve the accuracy of aggregate probabilistic forecasts (Atanasov et al., 2017; Budescu et al., 1997; Han & Budescu, 2022).

One potential pitfall in extremizing transformations is that they can excacerbate miscalibration in cases where the prior is biased. In many "wicked" forecasting problems, the majority is wrong (Prelec et al., 2017; Wilkening et al., 2022) and/or inaccurate forecasters express higher confidences (Koriat, 2008, 2012; Hertwig, 2012; Lee & Lee, 2017). In these cases, the average forecast often falls on the wrong side of 0.5. Extremizing wrong-sided average forecasts using the LLO transformation has the potential of pushing the forecast away from the true probability and can increase miscalibration rather than improve it.[2] Recent work by Lichtendahl Jr et al. (2022) shows that Bayesian aggregation often "antiextremizes" the average.

This paper explores a two-step algorithm that seeks to extremize the aggregate forecast while taking into account cases where the prior is biased and the majority may be wrong. We consider an environment in which individuals share a common prior that may be biased.[3] Forecasters receive independent signals conditional on the actual state such that the average probability forecast puts a higher probability on the actual state than the prior. When the prior is 0.5, the average forecast in these problems falls on the correct side of 0.5 as

[1] A simplified implementation sets $\delta = 1$ (Karmarkar, 1978; Erev et al., 1994; Shlomi & Wallsten, 2010), which is shown to improve calibration of the aggregate probability in forecasting geopolitical events (Mellers et al., 2014)

[2] Baron et al. (2014) discuss the issue of wrong-sided extremization in cases where the prior is 0.5. They consider an "extremize-then-aggregate", which can mitigate the issue in cases where wrong-sidedness is due to noise. As seen below, we concentrate on cases where the prior is biased and where all individual forecasters may be wrong-sided.

[3] We are agnostic as to where this bias might come from, but the setup is consistent with one where all forecasters initially observe the same common-signal and then receive a private idiosyncratic one. The common signal leads to the initial prior that differs from 0.5.

the overall crowd size grows large. Thus, in these cases extremization away from 0.5 can improve calibration. However, in a biased decision problem, wrong sidedness can occur. For example, if the prior is 0.7, there exists cases where the posterior is below 0.7 but above 0.5. In these cases, the LLO transformation would extremize the average forecast towards 1 which is contrary to the information contained in the forecaster's private signals.

We conjecture that a more appropriate extremization approach would be to extremize the data starting from the common prior rather than assuming that the prior is 0.5. To do so, we elicit each forecasters' estimate on the average forecast of others (referred to as their meta-precition) as well as their probabilistic forecast. We show that the meta-prediction can be used to estimate the prior in our setting and then implement an LLO transformation that recalibrates away from the estimated prior rather than using a neutral prior of 0.5.

To evaluate how well our algorithm calibrates, we estimate calibration curves across a variety of decision problems related to general knowledge, sports, and the price of art works. We find that our algorithm generates improves calibration relative to a variety of alternative algorithms that have been explored in the literature. These include the minimal pivoting algorithm Palley & Soll (2019), the knowledge weighting mechanism (Palley & Satopää, 2023), the meta probability weighting algorithm (Martinie et al., 2020), and the surprising overshoot (SO) algorithm (Peker, 2022). Our algorithm also generates very low brier scores across decision problems, suggesting that it has very good accuracy characteristics overall.

This paper contributes to the emerging literature on forecast aggregation methods that rely on meta-beliefs (Prelec et al., 2017; Palley & Soll, 2019; Martinie et al., 2020; Wilkening et al., 2022; Palley & Satopää, 2023; Peker, 2022). Meta-probability weighting aims to use forecasters' meta-prediction as well as their prediction to deal with biased priors or shared information. Forecasters whose prediction and meta-prediction diverge receive higher weights in the subsequent weighted average of predictions (Martinie et al., 2020). Minimal pivoting adjusts the average predictions based on how much it differs from the average meta-prediction (Palley & Soll, 2019). The adjustment corrects for the shared-information bias in the aggregate resulting from forecasters' common information. Knowledge-weighting proposes a weighted aggregation to address the shared information problem in a similar formal framework (Palley & Satopää, 2023). Individual weights are determined based on the weighted crowd's accuracy in meta-predictions. The surprising overshoot (SO) algorithm shares the same framework as minimal pivoting and knowledge weighting (Peker, 2022). The difference in how meta-predictions and predictions are distributed around the average prediction is considered as a measure of the shared-information bias. The Surprisingly Popular (SP) and Surprisingly Confident (SC) algorithms produce outcome predictions (Prelec et al., 2017; Wilkening et al., 2022). Similar to the methods mentioned above, the SC algorithm relies on probabilistic predictions and meta-predictions. The algorithm picks the outcome that the crowd considered more likely than the forecasters estimated.

Our formal framework is similar to Wilkening et al. (2022) and Martinie et al. (2020). Unlike Palley & Soll (2019), Palley & Satopää (2023) and Peker (2022), miscalibration in the average prediction could be due to factors other than shared information. However, our treatment of the meta-prediction data is similar to the knowledge-weighting algorithm. Knowledge-weighting considers forecasters' estimate on the average prediction of others as a noisy estimate of their true meta-prediction given their information. We follow a similar approach and estimate the prior through fitting a meta-prediction function using reported meta-predictions.

The rest of this paper is organized as follows: Section 2 introduces the Bayesian framework. Sections 3 discusses the existence of wrong side average forecasts in biased decision problems. Section 4 develops a robust recalibration rule that utilizes meta-predictions. Section 5 provides empirical evidence from experimental prediction tasks. Section 6 provides an overview of our contribution and concludes.

2 Framework

Our framework is similar to Wilkening et al. (2022) and Martinie et al. (2020). We are interested in predicting the occurrence E of a binary event. Let $o \in \{G, B\}$ be the state of the world where G and B represent "Good" and "Bad" states respectively. The occurrence of the event is associated with state G. Nature determines the state with unknown probability $q = Pr(o = G)$. Thus, an outcome forecast of E with the associated probability forecast (also referred to as "confidence") q would be perfectly well-calibrated. An aggregator elicits and aggregates judgments from a crowd of N forecasters, who share a common prior probability $p(G)$ on q. Each forecaster k receives a signal σ_k from $S \equiv \{s_1, \ldots, s_m\} \cup \{s_\emptyset\}$. Without loss of generality signals are normalized so that $s_i := p(G|s_i)$ where $p(G|s_i)$ is forecaster k's posterior belief on the probability of the true state being G when $\sigma_k = s_i$. The uninformative signal satisfies $s_\emptyset := p(G)$. Let $p(s_i|o)$ denote the probability of a signal s_i in state o, satisfying $\sum_i p(s_i|o) = 1$ for each $o \in \{G, B\}$. The conditional distribution of signals is represented by a likelihood matrix $[Q_{oj}]_{2\times(m+1)}$. The first and second rows give the likelihoods of each signal in states G and B respectively. Thus, $Q_{Gi} = Q_{1i} \equiv p(s_i|G)$. Also let $\bar{s}_o = \sum_i s_i p(s_i|o)$ be the expected signal in state o.

Given a signal s_i such that $p(s_i|G) + p(s_i|B) > 0$, the posterior belief is given by

$$p(G|s_i) = \frac{p(G)p(s_i|G)}{p(G)p(s_i|G) + p(B)p(s_i|B)} = s_i$$

Each forecaster k is asked to report i) a *prediction* P_k on the probability of G and ii) a *meta-prediction* M_k on the average of others' predictions. Since E is associated with state G, a probability prediction is a statement of confidence on the realization of E. We will assume that all forecasters report their best estimate for prediction and meta-prediction, and it is common knowledge that they do so. Let $P(\sigma_k)$ denote the prediction function, where $P(\sigma_k) = P_k = p(T|\sigma_k) = \sigma_k$. Also let $\bar{P} = \frac{1}{N}\sum_k P_k$ be the average prediction. Forecaster k's meta-prediction is given by $M(\sigma_k) = M_k = E[\bar{P}|\sigma_k]$. For a given outcome state o, expected average prediction is given by $E[\bar{P}|o] = \sum_i s_i p(s_i|o) = \bar{s}_o$. Then, forecaster k's meta-prediction can be written as $M_k = \sigma_k E[\bar{P}|G] + (1 - \sigma_k)E[\bar{P}|B]$. Since $P_k = \sigma_k$, M_k is a function of forecaster k's prediction P_k. The signal densities $\{Q_{Gi}, Q_{Bi}\}$ and $\{\bar{s}_G, \bar{s}_B\}$ are common knowledge to the forecasters but unknown to the aggregator.

We consider decision problems where the average prediction tends to be *underconfident*, i.e. $1 > q > \bar{s}_o$ if $o = G$ and $\bar{s}_o > q > 0$ if $o = B$. In other words, the average probability is not as extreme as it should be. Extremizing transformations would on average improve calibration as the actual frequency of G (and the realization of E) over many events with similar probability predictions would match the predicted frequency. Following Wilkening et al. (2022), we also categorize decision problems according to the properties of signal structure.

Definition (Biasedness). *A decision problem is unbiased if $s_\emptyset = 0.5$ and biased if $s_\emptyset \neq 0.5$.*

Section 3 discusses how extremization relates to the bias of the decision problem. We will demonstrate a potential pitfall in extremizing away from 0.5 in biased problems and propose a novel recalibration method.

3 Extremization in unbiased and biased problems

Section 1 discussed the previous literature suggesting that extremizing transformations improve the calibration of aggregate forecasts. Consider the average prediction $\bar{P}$ in our framework and let $t(\bar{P})$ be the recalibrated probability following a transformation function $t : [0, 1] \to [0, 1]$. The LLO transformation in Equation 1 is an extremizing transformation that leads to either $t(\bar{P}) > \bar{P} > 0.5$ or $t(\bar{P}) < \bar{P} < 0.5$ for $\bar{P} \neq 0.5$. To identify situations where extremization would improve calibration, we define the following property for average predictions:

Definition (Wrong-sided average prediction). *Average prediction $\bar{P}$ is wrong-sided if i) $o = G$ and $\bar{P} < 0.5$ or, ii) $o = B$ and $\bar{P} > 0.5$.*

If the average prediction is wrong-sided, extremization away from 0.5 could introduce further miscalibration to an already underconfident average prediction. To illustrate, consider the case $o = G$ and $\bar{P} < 0.5$. We have $E[\bar{P}|G] = \bar{s}_G < q$ due to underconfidence. Thus, we expect $\bar{P}$ to be smaller than q. If a probability transformation t extremizes away from 0.5, then we will have $t(\bar{P}) < \bar{P} < 0.5$. Thus, the transformed probability is expected to be even more inaccurate than $\bar{P}$.

When would the average prediction be wrong sided? Theorem 1 specifies how the average prediction relates to the properties of the decision problem:

Theorem 1. *In an unbiased decision problem, $E[\bar{P}|o]$ is not wrong-sided for all $o \in \{G, B\}$.*

This result follows from $s_\emptyset = E[\bar{P}] = 0.5$. By the law of iterated expectations, $E[\bar{P}] = qE[\bar{P}|G] + (1-q)E[\bar{P}|B]$. Since $E[\bar{P}|G] > E[\bar{P}|B]$, it must be that $E[\bar{P}|G] > 0.5 > E[\bar{P}|B]$.

Theorem 1 implies that, in an unbiased decision problem, extremizing away from 0.5 is expected to transform the average probability in the right direction. Suppose instead we have a biased decision problem where $E[\bar{P}|G] > s_\emptyset > E[\bar{P}|B] > 0.5$. The average prediction is expected to be wrong-sided if $o = B$. Extremizing away from 0.5 is likely to recalibrate the average prediction towards 1. However, $E[\bar{P}|B] > q > 0$ if $o = B$, implying that such extremization pushes the average towards the wrong extreme. An average prediction $\bar{P} \in (0.5, s_\emptyset)$ should be transformed towards 0 instead. Vice versa is true in biased problems with $s_\emptyset < 0.5$. Thus, a transformation function that extremizes average predictions away from $s_\emptyset$ instead of 0.5 would avoid transformations in the wrong direction. However, note that $s_\emptyset$ is unknown to the aggregator. Based on the prediction data alone, transformation t cannot be a function of $s_\emptyset$. Section 4 demonstrates how meta-predictions can be used to develop a transformation rule that recovers $s_\emptyset$ and recalibrate the average prediction accordingly.

4 Robust recalibration

Section 3 established that the average prediction may fall on the wrong side of 0.5 in biased decision problems. Extremizing such average predictions away from 0.5 could exacerbate the underconfidence. Extremizing away from $s_\emptyset$ is robust to biasedness of the decision problem as even the wrong-sided averages are recalibrated in the correct direction. This section shows that $s_\emptyset$ can be estimated from meta-prediction data to design a robust recalibration rule.

Consider the meta-prediction function $M(P_k)$ of forecaster k. We have the following result:

Theorem 2. $M(s_\emptyset) = P(s_\emptyset) = s_\emptyset$*, i.e. forecaster k's meta-prediction is equal to her prediction at the prior.*

Theorem 2 claims that the prediction and meta-prediction functions intersect at $s_\emptyset$. To establish this result, consider the meta-prediction function $M(s_\emptyset) = E[\bar{P}|s_\emptyset]$. We can write $E[\bar{P}|s_\emptyset] = E[\bar{P}|G]p(G|s_\emptyset) + E[\bar{P}|B]p(B|s_\emptyset)$. Note that $p(o|s_\emptyset) = p(o)$, i.e. the posterior probability conditional on the uninformative signal is equal to the prior at $\sigma_k = s_\emptyset$. Thus, we get

$$M(s_\emptyset) = E[\bar{P}|G]\, p(G) + E[\bar{P}|B]\, p(B). \tag{3}$$

Consider now the prediction function. We can write the expected signal of a forecaster k given the prior as follows:

$$\begin{aligned} E[\sigma_k|s_\emptyset] &= \sum_i p(G|s_i)p(s_i), && \text{(4a)} \\ &= \frac{Q_{Gi}\, p(G)}{\sum\limits_{o\in\{G,B\}} Q_{oi}p(o)} \sum_{o\in\{G,B\}} Q_{oi}\, p(o), && \text{(4b)} \\ &= p(G)\sum_i Q_{Gi} = p(G) = s_\emptyset. && \text{(4c)} \end{aligned}$$

Equation 4a follows from $s_i := p(G|s_i)$ and 4b rewrites the terms using the signal likelihoods. We have $\sum_i Q_{Gi} = 1$ in 4c as Q_{Gj} represents a probability density over all signals in state G. Thus, $E[\sigma_k|s_\emptyset]$ reduces to the prior $p(G)$, which is equal to the uninformative signal $s_\emptyset$. Note that we also have $E[\sigma_k|s_\emptyset] = \sum_i \sum_o s_i Q_{oi}$. Intuitively, Equation 4 establishes that $s_\emptyset$ is a weighted average of all posteriors (=signals) and states. We can write

$$\begin{aligned} s_\emptyset &= \sum_i s_i Q_{Gi} p(G) + \sum_i s_i Q_{Bi} p(B) \\ &= E[\bar{P}|G]p(G) + E[\bar{P}|B]p(B). \end{aligned} \tag{5}$$

Equations 3 and 5 imply that $M(s_\emptyset) = P(\emptyset) = s_\emptyset$. In other words, the prediction and meta-prediction functions intersect at the prior. Figure 1 provides an illustration:

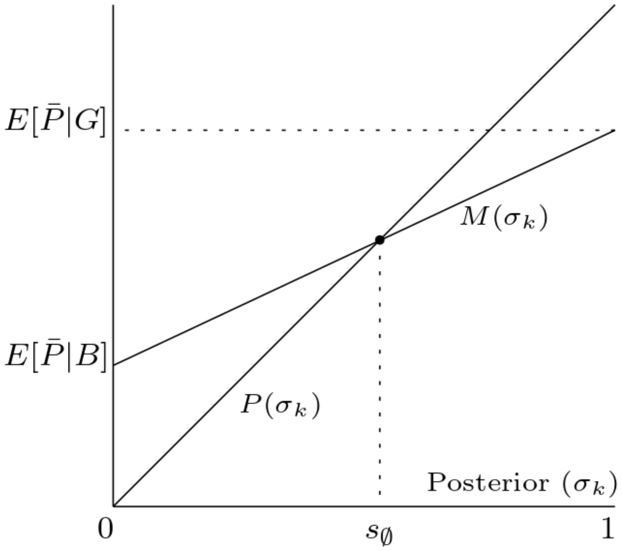


Figure 1: Exampe prediction and meta-prediction functions. THIS FIG IS JUST A PLACEHOLDER, CAN BE REPLACED WITH STH PRETTIER.

The prediction function coincides with the 45-degree line, while the meta-prediction function ranges between the expectations on $\bar{P}$ in each state. Both $P(\sigma_k)$ and $M(\sigma_k)$ are monotone increasing and linear functions in σ_k. Thus, $s_\emptyset$ is also the unique point of intersection.

Theorem 2 suggests that $s_\emptyset$ can be recovered if the parameters of the meta-prediction function are known. Solving for $s_\emptyset$ in $M(s_\emptyset) = P(s_\emptyset)$ gives

$$s_\emptyset = \frac{E(\bar{P}|B)}{1 - (E(\bar{P}|G) - E(\bar{P}|B))}$$

Since $\{E(\bar{P}|G), E(\bar{P}|B)\}$ are unknown to the aggregator, $s_\emptyset$ cannot be obtained directly. However, we can estimate $s_\emptyset$ using the prediction and meta-prediction data. Since $P_k = \sigma_k$, we can write $M(P_k) = E(\bar{P}|B) + (E(\bar{P}|G) - E(\bar{P}|B))P_k$. Consider the following linear regression model:

$$M_k = \beta_0 + \beta_1 P_k + \epsilon_k \tag{6}$$

Let $\{\hat{\beta}_0, \hat{\beta}_1\}$ denote the regression estimates. We consider $f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ as an estimate of the meta-prediction function M_k. Solving for $s_\emptyset$ in $f(s_\emptyset) = s_\emptyset$ we obtain an estimate for the prior, given by $\hat{s}_\emptyset = \hat{\beta}_0/(1 - \hat{\beta}_1)$ for $\hat{\beta}_1 \neq 1$.

Using the estimated prior $\hat{s}_\emptyset$, we propose a transformation function $t_r(\bar{P})$ that satisfies the following:

$$log\left(\frac{t_r(\bar{P})}{1 - t(\bar{P})}\right) = \gamma log\left(\frac{\bar{P}}{1 - \bar{P}}\right) + \gamma\left[log\left(\frac{\bar{P}}{1 - \bar{P}}\right) - log\left(\frac{\hat{s}_\emptyset}{1 - \hat{s}_\emptyset}\right)\right]. \tag{7}$$

Equation 7 suggests a linear transformation in log odds where, for $\gamma > 1$, $\bar{P} \geq s_\emptyset$ ($\bar{P} < s_\emptyset$) is adjusted towards 1 (0). Note that for $\hat{s}_\emptyset = 0.5$, Equation 7 is the same as Equation 2 with a reparametrization of slope as $1 + \gamma$ instead of γ and intercept being zero. In the special case of the estimated prior being unbiased ($\hat{s}_\emptyset = 0.5$), t_r reduces to the LLO transformation away from 0.5 with $\delta = 1$, also known as the Karmarkar equation (Karmarkar, 1978). Solving Equation 7 for $t_r(\bar{P})$, we get

$$t_r(\bar{P}) = \frac{\delta \bar{P}^{1+\gamma}}{\delta \bar{P}^{1+\gamma} + (1 - \bar{P})^{1+\gamma}} \tag{8}$$

where $\delta = [(1 - \hat{s}_\emptyset)/\hat{s}_\emptyset]^\gamma$. Unlike simple extremization away from 0.5, $t_r(\bar{P})$ is robust to wrong-side average predictions. The average is transformed away from $\hat{s}_\emptyset$ instead of 0.5. If $\hat{s}_\emptyset$ estimates the unknown $s_\emptyset$ accurately, we should expect t_r to adjust wrong-sided average predictions in the correct direction.

Section 5 tests the robust recalibration rule $t_r(\bar{P})$ using experimental data sets. Note that the case of $\hat{s}_\emptyset = 0.5$ (Karmarkar equation) corresponds to the extremizing transformation proposed by Baron et al. (2014). Their LLO extremization can be considered as an implementation of t_r where all decision problems are considered unbiased. Thus, we will consider $t_r(\bar{P})$ with $\hat{s}_\emptyset = 0.5$ in all problems as a benchmark that represents "always extremize away from 0.5". This benchmark allows us to evaluate if the use of meta-predictions to estimate $s_\emptyset$ improves the calibration. The analysis will compare t_r with various aggregation mechanisms that generate probability forecasts.

5 Empirical evidence

This section presents empirical evidence for the effectiveness of robust recalibration. We use data from experimental prediction tasks where subjects are asked to report a meta-prediction as well as their prediction. Section 5.1 introduces the data sets. Section 5.2 presents preliminary evidence on the existence of wrong-sided average predictions and discusses estimated priors. Section 5.3 offers a comparative analysis on the calibration of transformed probabilities.

5.1 Data Sets

We investigate the empirical performance of robust recalibration using four distinct types of experimental tasks. In the first, subjects are presented with simple scientific statements. For each statement, they report a probabilistic prediction on the statement being true as well as a meta-prediction on the average of other subjects' predictions. Martinie et al. (2020)

and Wilkening et al. (2022) collect data from 500 such statements. We also use data from subsequent replications on a subset of these statements. Each implementation recruits a new sample of subjects. Thus, we treat each statement-forecasting crowd combination as a different forecasting task. The resulting 'Science' data set includes 680 tasks in total and the number of subjects in a task varies between 79 and 98.

The second data set, referred to as 'States' data, is also collected by Wilkening et al. (2022). The States data set includes 50 tasks. Each task presents a statement on the largest city of a U.S. state being the capital city of the corresponding state. A total of 89 subjects report probabilistic predictions and meta-predictions on the truth of each statement.

Our final data source is Howe et al. (2023), who collect predictions and meta-predictions on various domains. The art evaluation tasks elicit judgments on the prices of artworks. Subjects see a picture of a drawing and expected to predict how likely it is that the market value is more than $10000. The 'Artwork' data set includes 40 such items, implemented in two replications to produce 80 tasks. The sample size for each task varies between 79 and 87 subjects. Finally, the 'NFL domain' tasks present 50 trivia statements about the NFL draft to a US-based subject pool. Similar to the Artwork data, two runs produce 100 tasks in total. Appendix A provides a sample of tasks from each data set. We have a grand total of 910 tasks in our data.

5.2 Preliminary evidence on priors and wrong-sided average predictions

Robust recalibration is expected to improve over simple extremization in transforming wrong-sided average probabilities. Since the correct answer in our prediction tasks are known, we can first investigate the frequency of wrong-sided averages. Figure 2 shows the number of tasks in each data set where the average prediction is wrong-sided.

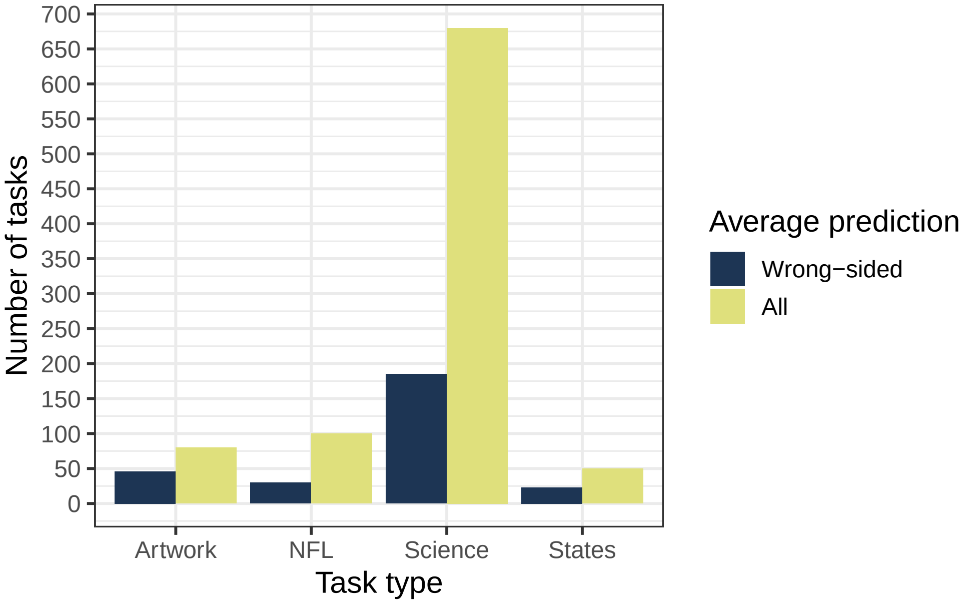


Figure 2: Wrong-sided averages in each data set

Figure 2 demonstrates that the average prediction is wrong-sided in a considerable number of tasks in each data set. We should expect better calibration in transformed probabilities produced by robust recalibration. Figure B2 in Appendix B shows the distribution of average predictions in "True" and "False" statements. Wrong-sided average predictions are especially common in "False" statements, suggesting that the crowd may often have an upward bias in their prior on the likelihood of "True". Figure 3 provides evidence along these lines.

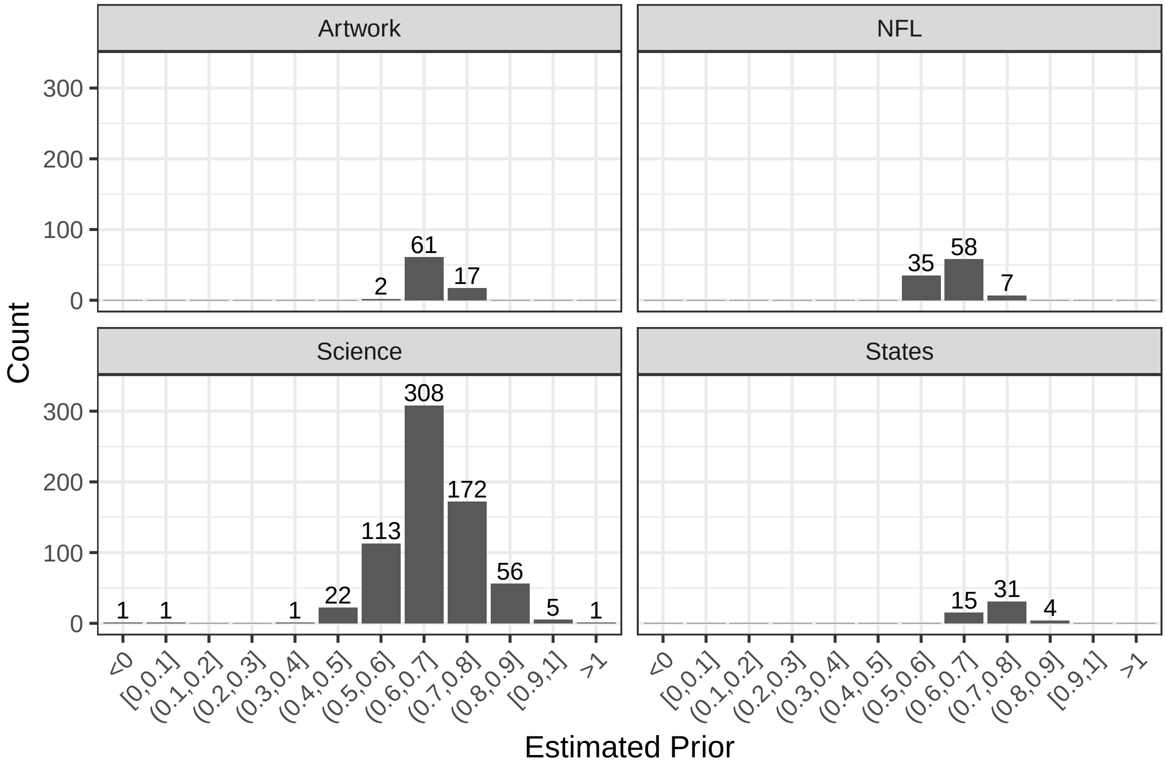


Figure 3: The distribution of estimated priors in each data set.

Estimated priors are typically higher than 0.5, which would predict a higher frequency of wrong-sided averages among "False" statements. Robust recalibration will often transform an average prediction above 0.5 towards 0 while extremization pushes the same average further towards 1. We should note that in two tasks of the Science data, the estimated priors lie outside $(0, 1)$. This can be considered as a failure to estimate $s_\emptyset$ accurately. In our implementation, we set $\hat{s}_\emptyset = 0.5$ in the two instances where $\hat{s}_\emptyset \notin (0, 1)$. In other words, robust recalibration reverts to simple extremization when $\hat{s}_\emptyset \notin (0, 1)$.

Tables 1a and 1b show how average predictions compare to 0.5 and estimated priors respectively.

(a)

	Correct answer		
	True	False	Total
$\bar{P} > 0.5$	416	263	679
$\bar{P} < 0.5$	21	210	231
Total	437	473	910

(b)

	Correct answer		
	True	False	Total
$\bar{P} > \hat{s}_\emptyset$	270	40	310
$\bar{P} < \hat{s}_\emptyset$	167	433	600
Total	437	473	910

Table 1: Average prediction vs. 0.5 or estimated prior for "True" and "False" statements

Table 1a confirms that wrong-sided average predictions are more common among the "False" statements in our data sets. The average prediction is above 0.5 in 263 of 473 "False" statements. Extremization incorrectly transforms these average probabilities towards 1. Figure 3 showed that most estimated priors are above 0.5. As a result, robust recalibration correctly transforms 433 of 473 "False" statement towards 0 instead of 1. However, estimated priors do not always suggest the correct direction for transforming the average prediction. A binary classification (i.e. a probabilistic forecast of 0 or 1) based on the relationship between the average prediction and the estimated prior would especially misidentify many "True" statements as "False". Section 5.3 implements extremization, robust recalibration and various aggregation algorithms in our data sets to investigate the calibration of the resulting probabilities.

5.3 Results

This section investigates the accuracy and calibration of the robust-recalibrated probability forecasts. We will run comparative analyses where alternative methods are implemented as benchmarks. The first analysis compares robust recalibration to the average prediction and the average extremized away from 0.5. The former is the untransformed simple average of predictions while the latter transforms the average prediction using Equation 8 with $\hat{s}_\emptyset = 0.5$, which corresponds to $\delta = 1$. We consider $\gamma \in \{1, 1.5, 2, 2.5\}$ in our implementations of Equation 8 for both extremization and robust recalibration. Then we implement vaious alternative aggregation algorithms from recent literature that use meta-predictions to improve accuracy. More information on these algorithms are given below. Aggregate predictions are calculated for all 910 tasks.

Figure 4 shows the distribution of Brier scores of the average prediction, extremized average and robust-recalibrated prediction across all tasks. Lower scores indicate more accurate forecasts. Each row in the 4x3 grid shows the implementation of extremization away from 0.5 and robust recalibration for various values of γ. We also classify the tasks in terms of how extreme the untransformed average prediction is. Average probability predictions above 0.5 correspond to the confidence for T, while for an average probability below 0.5, one minus the probability gives the confidence for F. The coloring in Figure 4 breaks down the distribution of score for five different confidence levels of the corresponding average prediction.

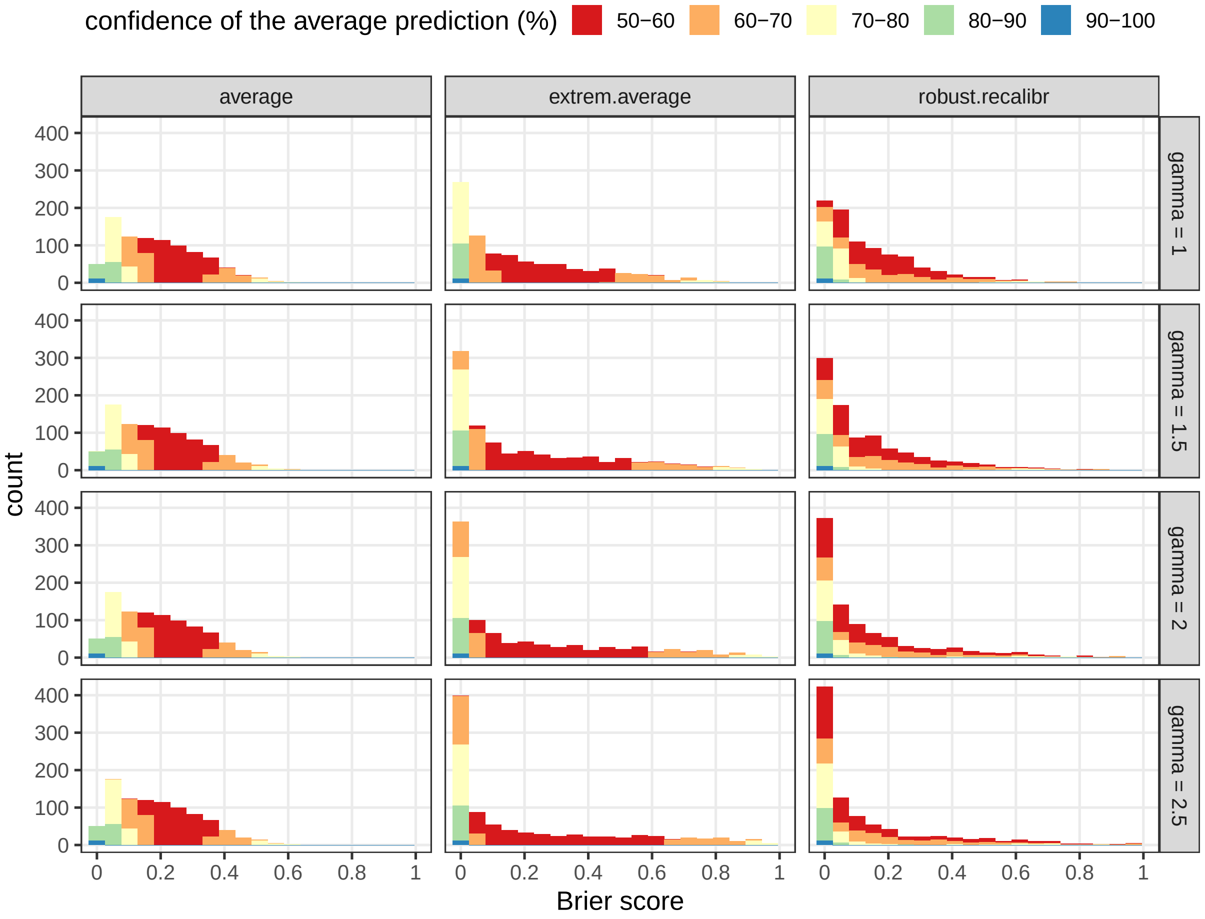


Figure 4: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

Figure 4 suggests that extremizing the average prediction away from 0.5 increases the expected accuracy. This result agrees with previous findings on extremization (Han & Budescu, 2022). Similarly, robust recalibration produces more accurate probability predictions than the average prediction. Figure B1 in Appendix B indicates even lower Brier scores than extremization in a pairwise comparison (also see Table B1 for a Wilcoxon test of median scores). However, the scores for different levels confidence of the average prediction suggest another pattern. Robust recalibration is particularly effective in transforming low-confidence average predictions. Figure B1 confirms that robust recalibration achieves better Brier scores than extremization in such tasks. Recall that wrong-sided averages occur mostly in "False" statements in our experimental prediction tasks (Table 1), as predicted by the estimated priors (Figure 3). Figure B2 in Appendix B indicates that most such wrong-sided averages are within $(0.5, 65)$, thus classifying them as predictions of "True" with low confidence. Extremization wrongly transforms these average predictions into high-confidence "True" predictions. Robust recalibration pushes the average prediction away from the estimated prior instead. Since estimated priors are often higher than 0.5, robust recalibration accurately classified many wrong-sided averages as predictions of "False" with low confidence and transformed them towards 0, which produces better Brier scores.

Note that many average predictions are not wrong-sided and extremizing away from 0.5 improves accuracy in such tasks. However, we may expect the extremization of wrong-sided averages to distort the calibration, in particular when wrong-sidedness is more prevalent for one of the states, as is the case with "False" in our prediction tasks. Untransformed average predictions are also likely to be miscalibrated due to underconfidence. Robust recalibration corrects for the underconfidence while avoiding the extremization of wrong-sided averages. Thus, we expect robust-recalibrated probabilities to more accurately reflect the actual frequencies. Figure 5 depicts the calibration curves for simple average, extremized average and the transformed probability from robust recalibration. Each panel represents extremization and robust recalibration for a value of γ where simple averages are the same for different γ. Probability predictions from each method are categorized in bins $\{[0, 0.1], (0.1, 0.2], \ldots, (0.9, 1]\}$. The proximity of the predicted probability of "True" and the actual proportion of "True" in each bin provides a measure of calibration. The shaded regions represent the range of proportion "True" at which the probability predictions in the corresponding bin are considered well-calibrated. Intuitively, the shaded regions are analogous to the 45-degree line of perfect calibration.

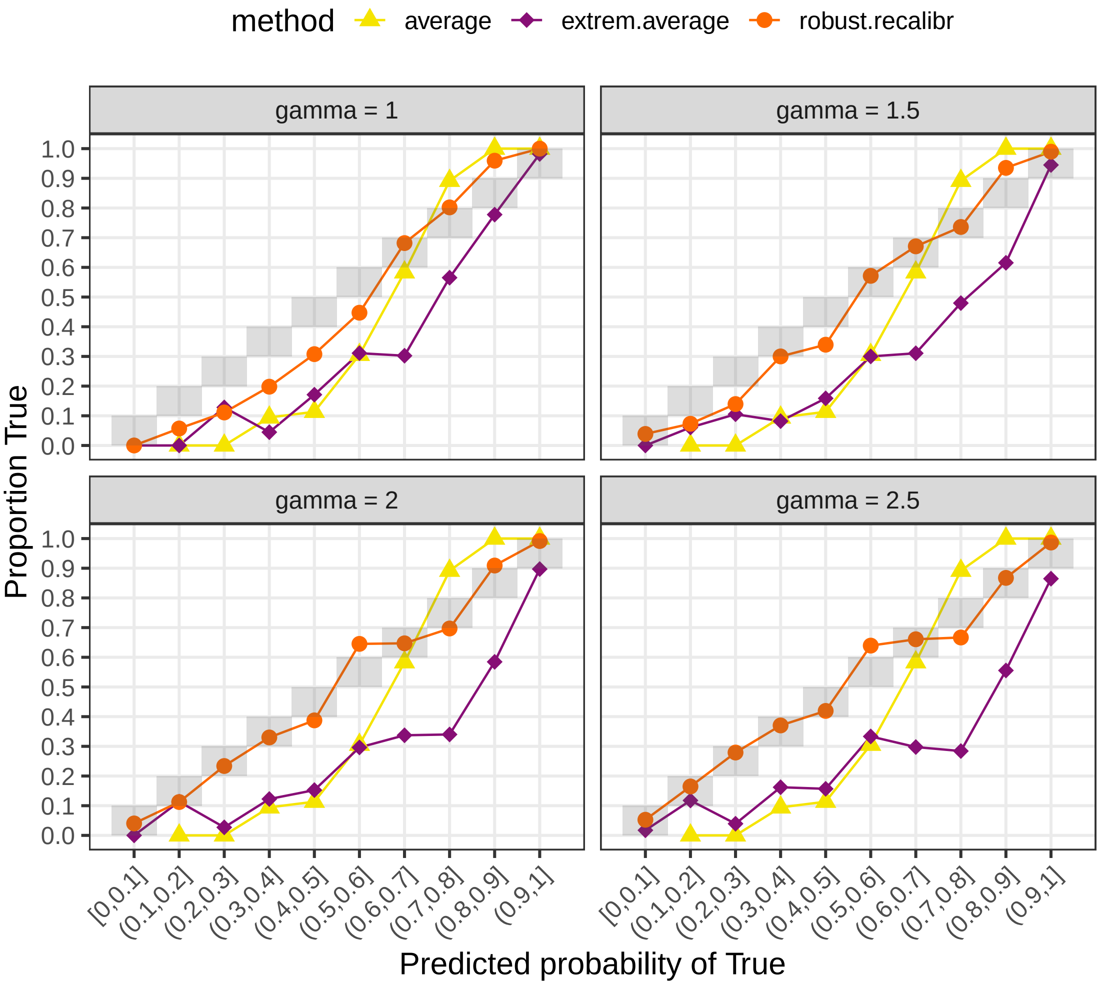


Figure 5: Calibration curves for simple average, extremized average and robust-recalibrated probabilities.

Figure 5 suggests that the transformed probabilities from robust recalibration achieve better calibration. In particular for $\gamma \in \{2, 2.5\}$, robust-recalibrated probabilities on "True" closely reflect the actual frequency of "True" in most bins. In contrast, for extremized averages, the actual proportion of "True" is typically lower than the predicted probability in the corresponding bin. In other words, extremized averages typically overestimate the probability of "True". Figures 4 and 5 together imply that the robust recalibration presents a probability transformation that manages to improve both accuracy and calibration.

Figure 5 compares robust recalibration to simple average and extremization. Unlike robust recalibration, these benchmarks do not require meta-prediction data. The use of meta-beliefs allows implementation of more advanced aggregation algorithms that can transform wrong-side predictions in the correct direction. We now elaborate on these algorithms and implement them as benchmarks for the second comparative analysis.

Meta-probability weighting constructs a weighted average of probabilistic forecasts, where a forecaster's weight is proportional to the absolute difference between her prediction and meta-prediction (Martinie et al., 2020). Consider the scenario where the average forecast is wrong-sided because only a minority of forecasters endorse the correct state. If accurate forecasters anticipate that they are in the minority, we may observe a larger absolute difference between their own forecast and meta-prediction on the average forecast of others. In that case, such forecasters would be weighted more heavily, potentially transforming a wrong-sided forecast correctly in the opposite direction of extremization.

Knowledge-weighting proposes another weighted averaging scheme (Palley & Satopää, 2023). The optimal weights minimize the "peer-prediction gap", which measures the difference between a weighted average of forecasters meta-predictions and the actual realization of the average forecast. If forecasters use their information optimally in forming meta-predictions, the weights that minimize the peer-prediction gap minimize the error in aggregate forecast as well. Intuitively, if the accurate minority of forecasters are also more accurate in their meta-predictions, knowledge-weighting is expected to put a higher weight on their forecasts, which may transform a wrong-sided average forecast in the correct direction. Knowledge-weighting is applicable in all forms of continuous variables, including non-probabilistic predictions. The knowledge-weighted prediction was outside of $[0, 1]$ in some of our tasks. We winsorize these predictions such that aggregates below 0 (above 1) are set at 0 (1).

Minimal pivoting uses meta-prediction data to correct for a potential shared-information bias in the average forecast (Palley & Soll, 2019). Information commonly available to forecasters may bias probabilistic forecasts in a particular direction, which could lead to a wrong-side average forecast. Minimal pivoting adjusts the average forecast according to the difference between average forecast and the average meta-prediction. Meta-predictions are expected to be influenced more heavily by the shared information because forecasters anticipate that their peers will also incorporate it in their forecasts. The pivoting procedure moves the average away from the shared information. The correction for the shared-information bias may improve the calibration as well. Similar to the knowledge-weighting algorithm, transformed probabilities that are outside of $[0, 1]$ are winsorized.

The Surprising Overshoot (SO) algorithm is another aggregation method that addresses the shared-information problem (Peker, 2022). Information available to a forecaster determines the meta-prediction as well as the prediction, resulting in a positive correlation between the two. Then, prediction and meta-prediction of an individual should typically fall on the same side of a well-calibrated average prediction. As mentioned above, shared information biases meta-predictions more strongly. A significant difference between the percentage of predictions and meta-predictions that overshoot the average prediction would constitute an "overshoot surprise", which suggests a miscalibration in the average prediction itself. The SO algorithm produces an aggregate forecast that corrects for the shared-information bias using the information in the size and direction of an overshoot surprise.

Figure 6 presents the frequency distribution of Brier scores. We set $\gamma = 2$ for the implementation of robust recalibration. Similar to Figure 4, color coding indicates the confidence level of the average prediction in the corresponding prediction task.

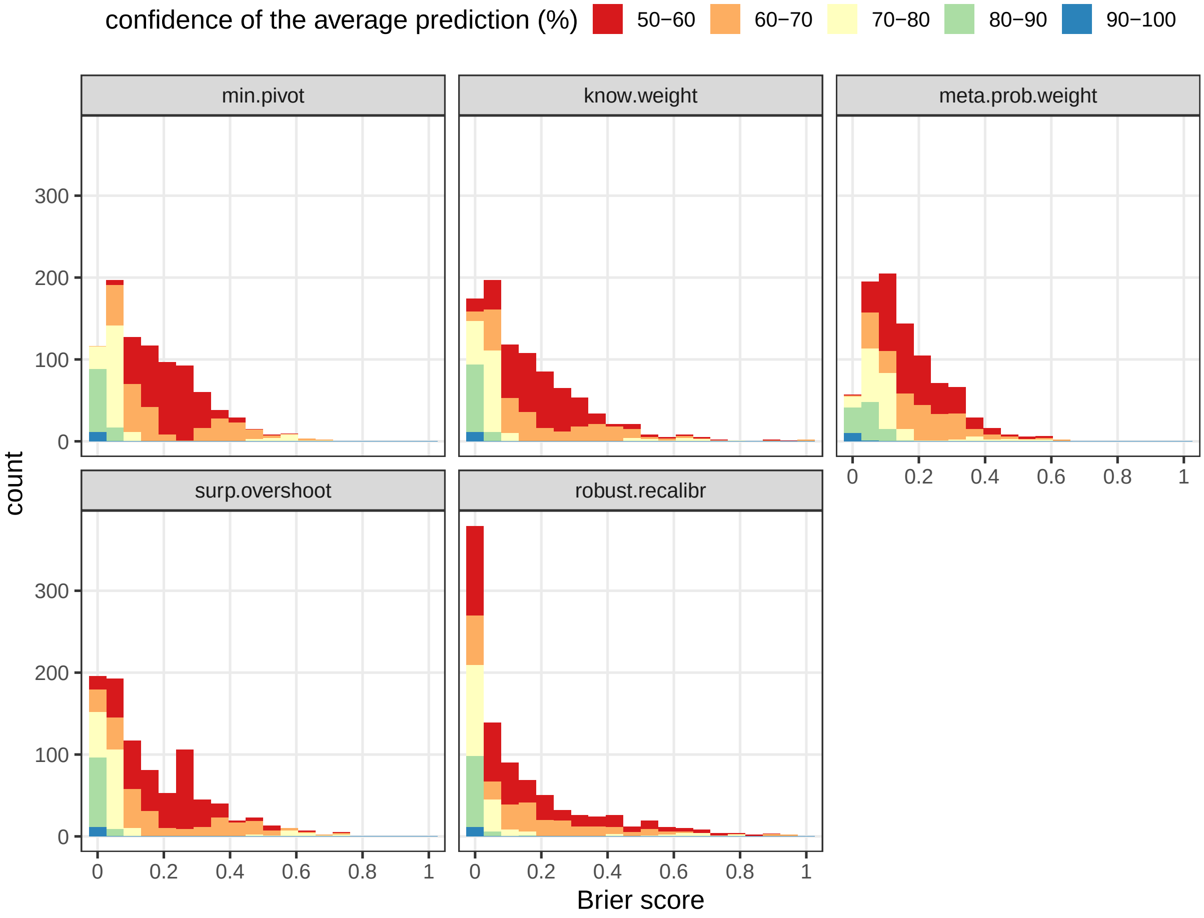


Figure 6: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

Robust recalibration achieves very small Brier scores more often than the benchmarks. The difference between the Brier scores of algorithms is significant (ANOVA test, F-value = 5.874, p = 0.000103). Pairwise comparisons suggest that robust recalibration achieves the highest accuracy [4]. The aggregation algorithms in consideration need not have an extremizing effect. The success of robust recalibration in achieving high Brier scores can be explained by its ability to push an underconfident forecast towards the appropriate extreme. Figure 7 presents the calibration curve of each benchmark and robust recalibration with $\gamma = 2$. Similar to Figure 5, predictions are categorized in bins and shaded regions represent the implied

[4] Tukey HSD test on mean difference: mean diff = -0.0322 vs min.pivot, $p < 0.0001$; mean diff = -0.0256 vs know.weight, $p = 0.0031$; mean diff = -0.0238 vs meta.prob.weight, $p = 0.0076$; mean diff = -0.0224, $p = 0.0146$.

proportion of “True”. Each dots represents the realized proportion of “True” for the tasks an which the method’s probability prediction lies in the corresponding bin. A calibration curve closer to the 45-degree line would indicate proper calibration for the predictions of the associated method.

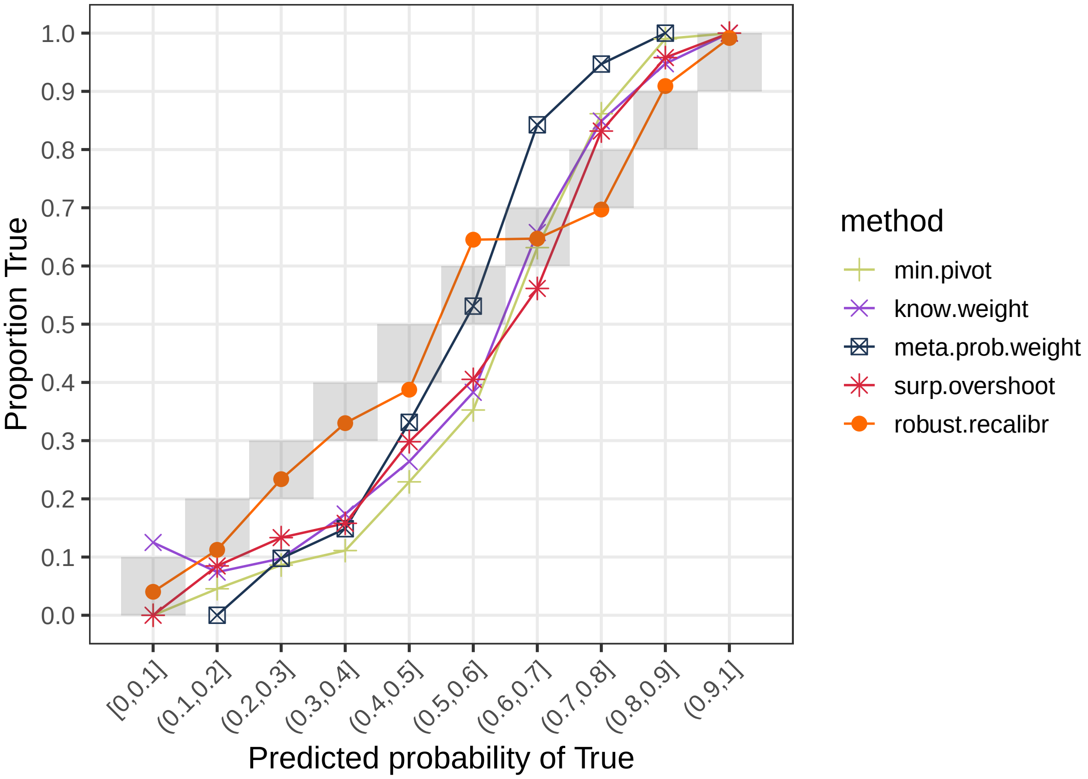


Figure 7: Calibration curves for simple avarage, extremized average and robust-recalibrated probabilities.

For predicted probabilities within $(0.5, 0.6]$, meta-probability weighting produces the most well-calibrated probabilities. Robust recalibration achieves the best calibration among alternatives if the whole range of probabilities is considered.

6 Conclusion

Probabilistic forecasts are often too conservative, which leads to average probability forecasts not being sufficiently extreme. Previous work documented that extremizing transformations that adjust the average away from 0.5 improve calibration. However, such transformations may have shortcomings. In some forecasting problems, the crowd may have a biased prior that favors a certain outcome. Then, the average forecast may put a higher probability on the wrong outcome even when individuals receive informative signals conditional on the correct outcome. Extremizing a wrong-sided average forecast would introduce further miscalibration.

This paper proposes a probability transformation that is robust to biased priors. We show that forecasters' meta-beliefs on others' predictions can be used to estimate the prior. Then, we propose a recalibration function that transforms the average away from the estimated prior instead of 0.5. A bias in crowd's prior probability is reflected in the estimated prior. Thus, unlike simple extremization away from 0.5, robust recalibration is capable of correctly transforming wrong-side averages in the opposite direction of extremization. Evidence from four distinct experimental tasks suggest that robust recalibration is effective in improving the calibration of probability forecasts. Robust-recalibrated probabilities predict the actual frequency of occurrence more accurately than extremized averages as well as the forecasts from advanced aggregation algorithms that rely on meta-beliefs.

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., ... Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130.

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

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment. part ii: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*(3), 173–188.

Camerer, C. F., & Ho, T.-H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of risk and uncertainty*, *8*(2), 167–196.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological review*, *101*(3), 519.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human perception and performance*, *3*(4), 552.

Han, Y., & Budescu, D. V. (2022). Recalibrating probabilistic forecasts to improve their accuracy. *Judgment and Decision Making*, *17*(1), 91.

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, *336*(6079), 303–304.

Howe, P. D., Martinie, M., & Wilkening, T. (2023). Using cross-domain expertise to aggregate forecasts when within-domain expertise is unknown. *Decision*.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational behavior and human performance*, *21*(1), 61–72.

Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, *52*(1), 111–127.

Lee, M. D., & Lee, M. N. (2017). The relationship between crowd majority and accuracy for binary decisions. *Judgment & Decision Making*, *12*(4).

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., Jose, V. R., & Winkler, R. L. (2022). Extremizing and antiextremizing in bayesian ensembles of binary-event forecasts. *Operations Research*, *70*(5), 2998–3014.

Martinie, M., Wilkening, T., & Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *Plos one*, *15*(4), e0232058.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … others (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106–1115.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Palley, A., & Satopää, V. A. (2023). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Management Science*.

Palley, A., & Soll, J. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309.

Peker, C. (2022). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*. doi: 10.1007/s11238-022-09899-4

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

Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors' probability estimates. *Psychonomic bulletin & review*, *17*(4), 492–498.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* New York, NY, US: Doubleday & Co.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine learning*, *95*(3), 261–289.

Wilkening, T., Martinie, M., & Howe, P. D. (2022). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*, *68*(1), 487–508.

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management science*, *42*(12), 1676–1690.

Appendices

A Prediction tasks

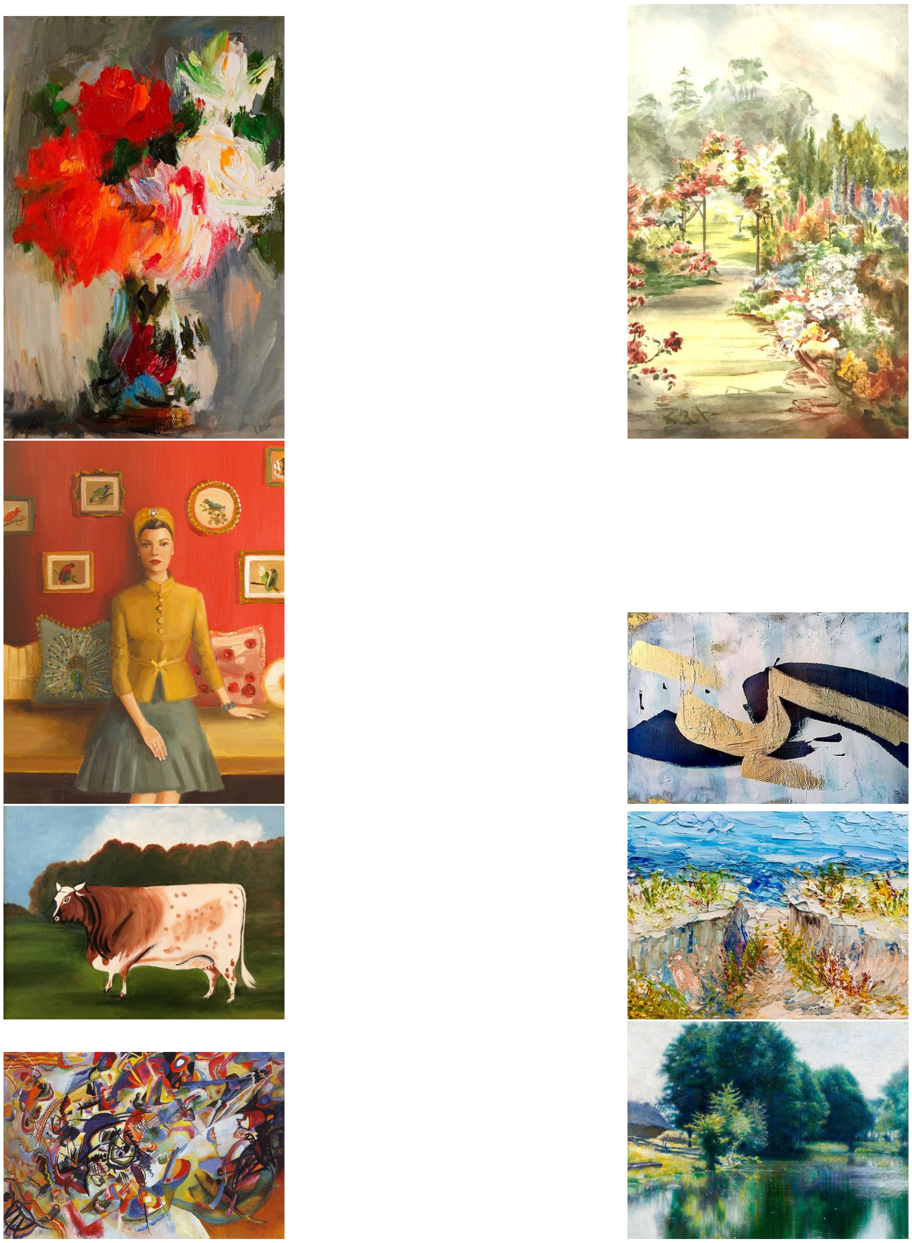
Table A1: Sample statements from Science and States data. See the supplemental material of Wilkening et al. (2022) for full list of statements

Data set	Statement
Science	Scurvy and anemia are diseases not caused by bacteria or viruses
Science	Secondary industries dominate the market in emerging economies
Science	Earthquakes and volcanoes typically occur at the boundaries of tectonic plates
Science	A substance with a pH of 8 is a strong acid
Science	Hamsters hate to run
Science	Plant cells are easier to clone than animal cells
Science	Convex lenses are used to correct for short-sightedness
Science	Darwin's theory was not widely accepted when it was first published in the late 19th century
Science	Increasing the number of impermeable rocks in rivers help decrease the flood risk
States	Jacksonville is the capital city of Florida
States	Los Angeles is the capital city of California
States	Denver is the capital city of Colorado

Table A2: Sample NFL statements

Statement
In the 2018 NFL draft, Mark Andrews was drafted by the Minnesota Vikings
In the 2018 NFL draft, the New York Giants were the only team to draft a player out of FCS champion North Dakota State University
In the 2017 NFL draft, the Big Ten was one of the athletic conferences where no players were drafted that year
In the 2016 NFL draft, Rico Gathers was drafted by the Oakland Raiders
In the 2016 NFL draft, David Onyemata was drafted by the New Orleans Saints
In NFL rules, a player who wears illegal equipment is to be suspended for the next two games
In NFL rules, a delay of game penalty at the start of either half is a 5-yard penalty
In NFL rules, the penalty for attempting to use more than 3 timeouts in a half is 5 yards
In NFL, a ”Hail Mary” is a play in which the receivers are all sent downfield towards the end zone
In NFL, a ”two-point conversion” is a play a team attempts instead of kicking a one-point conversion immediately after it scores a touchdown

Figure A1: Sample items from the Artwork data set



B Additional figures

Table B1: Paired Wilcoxon signed rank test of Brier scores, Robust recalibration vs Extremizing away from 0.5

Gamma	Test stat.	p-value
1	$V = 143280$	$p < 0.0001$
1.5	$V = 148088$	$p < 0.0001$
2	$V = 151761$	$p < 0.0001$
2.5	$V = 154699$	$p < 0.0001$

Figure B1: Pairwise differences in Brier score, Robust recalibration vs extremized average for $\gamma \in \{1, 1.5, 2, 2.5\}$. The total number of observations is 910.

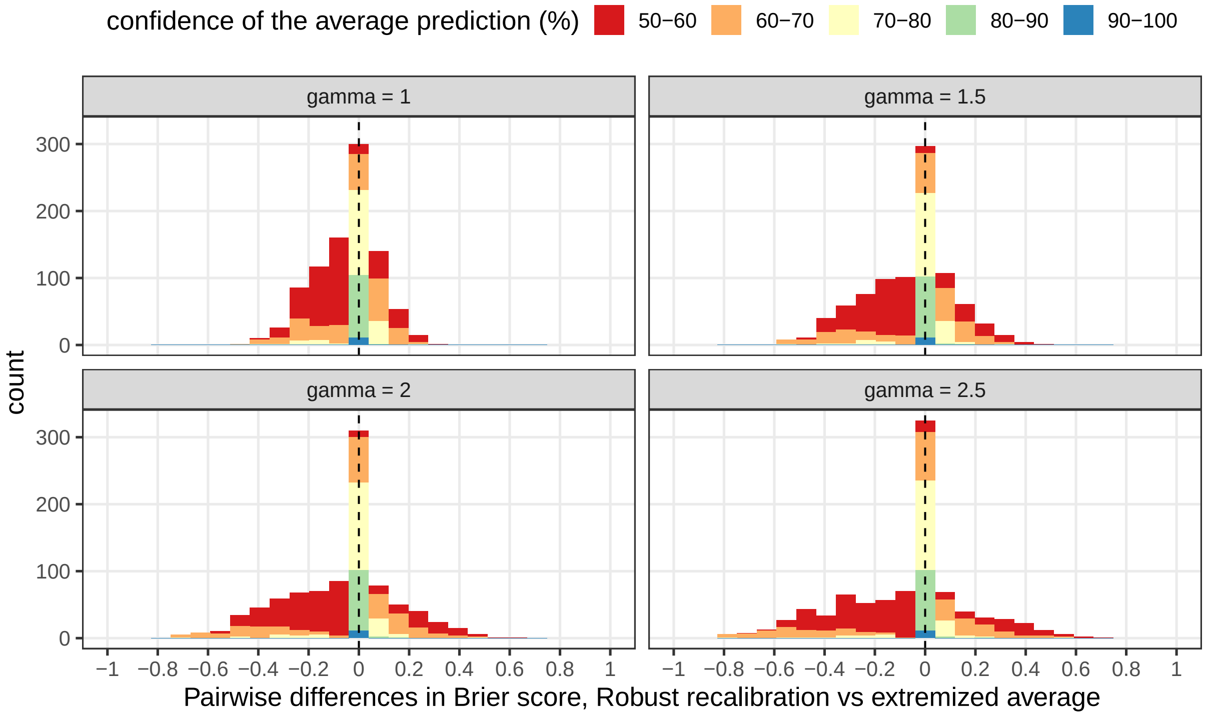
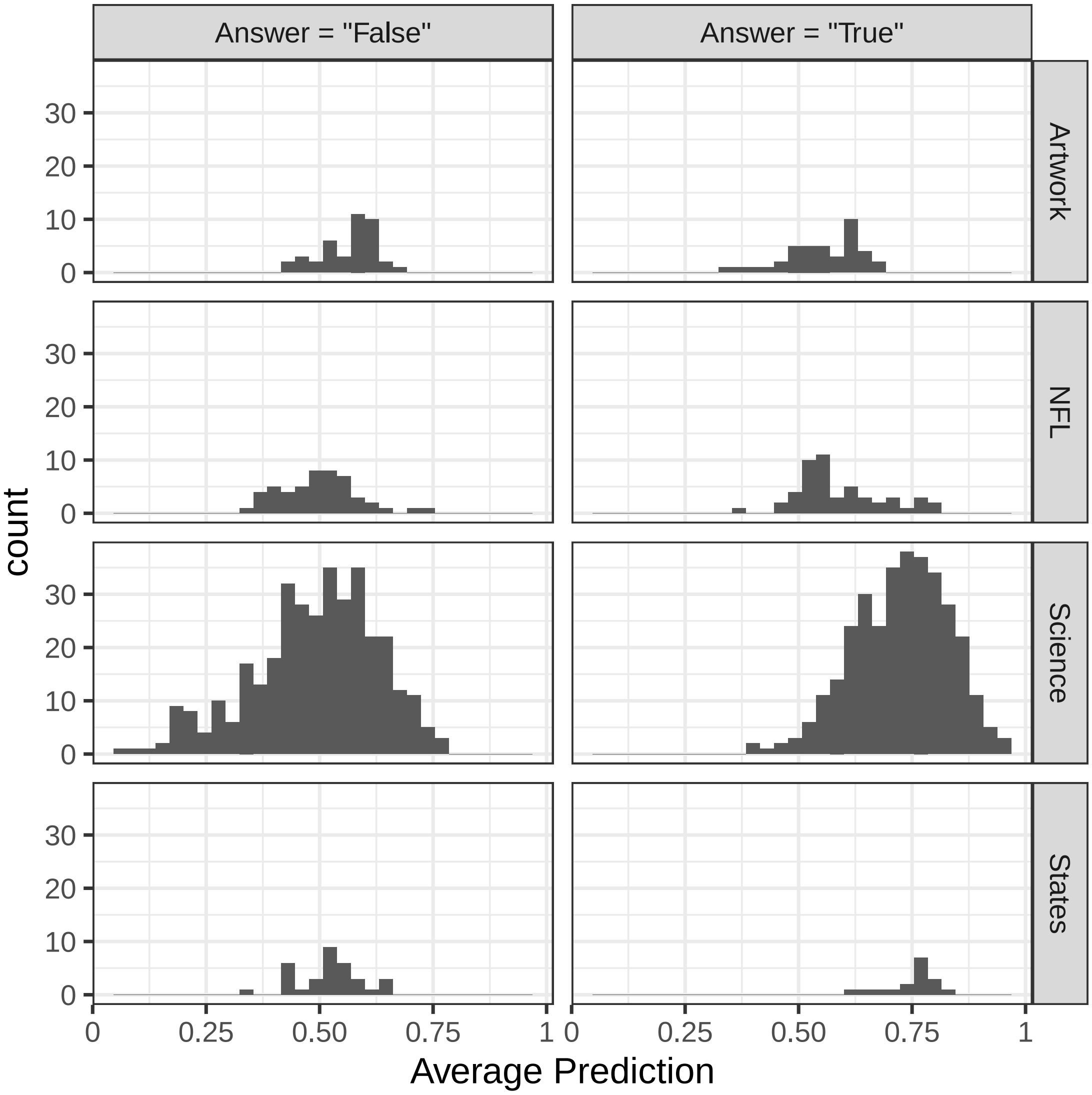


Figure B2: The distribution of average predictions for “True” and “False” statements in each data set.


Answer = "False"
Answer = "True"
Artwork
NFL
Science
States
count
Average Prediction