Robust recalibration of aggregate probability forecasts using meta-beliefs*

Cem Peker†[1] and Tom Wilkening[2]

[1]Divison of Social Science, New York University Abu Dhabi

[2]Department of Economics, Universiy of Melbourne

October, 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.

Keywords— judgment aggregation, wisdom of crowds, forecasting, extremization, recalibration, meta-beliefs

*We thank the audiences at 2022 INFORMS Annual Meeting and European Decision Sciences Seminar for helpful comments. Cem Peker gratefully acknowledges financial support from the NYUAD Center for Behavioral Institutional Design (C-BID), funded by Tamkeen under the NYUAD Research Institute Award CG005. Tom Wilkening gratefully acknowledges financial support from the Australian Research Council (Future Fellowship Research Grant, FT190100630).

†**E-mail addresses:** cem.peker@nyu.edu (C. Peker), tom.wilkening@unimelb.edu.au (T. Wilkening).

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

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

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 improving accuracy.[2]

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 an event may occur, which 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 that the event occurs is 0.5, the average forecast in these problems always falls on the correct side of 0.5 as the overall crowd size grows large. Thus, in these cases extremization away from 0.5 can improve calibration. However,

[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] The importance of using a value other than 0.5 as the basis for extremization is explored in Lichtendahl Jr et al. (2022), which shows that Bayesian aggregation often "antiextremizes" the average. 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.

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-predictions 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, 2023). 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 higher order beliefs (Prelec et al., 2017; Palley & Soll, 2019; Martinie et al., 2020; Wilkening et al., 2022; Palley & Satopää, 2023; Peker, 2023; Chen et al., 2021). The elicitation of higher-order beliefs allows the researcher additional information about the signals that individuals receive that can be useful in cases where signals are either correlated or noisy and where forecasters themselves have more information about the data-generating process than the aggregator.

Meta-prediciton algorithms have been developed both for binary classification (e.g., Prelec et al. (2017); Wilkening et al. (2022); Chen et al. (2021)) problems and in settings like

ours where the aggregator wishes to make a probabilistic forecast. In this second class of problems, four main alternative approaches have been proposed: meta-probability weighting, minimal pivoting, knowledge weighting, and surprising overshooting. 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 that seeks to overweight forecasters who are better at predicting the forecasters of their peers (Palley & Satopää, 2023). Finally, the surprising overshoot (SO) algorithm corrects for shared information using the observation that the prediction and meta-prediction of an individual should both fall on the same side of a well-calibrated average (Peker, 2023).

Our formal framework is similar to Wilkening et al. (2022) and Martinie et al. (2020) in that individuals receive private noisy signals but share a common biased prior. This framework naturally introduces conservative forecasts since all individals have only imperfect information about the true state. Palley & Soll (2019), Palley & Satopää (2023) and Peker (2023) use an alternative framework that allow for intermediate types of shared information but places stronger restrictions on the types of signals received. The framework used in knowledge weighting lies between the two approaches and considers an environment where forecasters' make noisy predictions and meta-predictions based on their true information.

Our recalibration procedure relies on a regression approach that is similar to the fitting technique used in Palley & Satopää (2023) that seeks to estimate a meta-prediction function using reported predictions and meta-predictions. Regression approaches have also been proposed by Libgober (2023) to identify information regarding the underlying data-generating process.

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 and develops the robust recalibration method that utilizes meta-predictions. Section 4 provides empirical evidence from experimental prediction tasks. Section 5 provides an overview of our contribution and concludes.

2 Framework

Our framework is similar to Wilkening et al. (2022) but adapted to the forecasting domain. We are interested in predicting the probability that a binary even E will occur. The probability that this event occurs varies with a state that is unobservable to the decision maker. However, forecasters receive signals regarding the underlying state and have common knowledge regarding the likelihood of each potential signal in each potential state.

The main text considers the ideal setting where there are two potential underlying states.[4] Let $\omega \in \{\omega_G, \omega_B\}$ be the state of the world where G and B represent "Good" and "Bad" states respectively. The occurrence of the event occurs with probability $Pr(E|\omega_G) = g$ in the good state and with probability $Pr(E|\omega_B) = b$ in the bad state, satisfying $g > b$. Nature determines the state with unknown probability $Pr(\omega = \omega_g)$. Thus, a probability forecast g of E when the state is good and b when the state is bad would be a perfectly well-calibrated forecast.

An aggregator elicits and aggregates judgments from a crowd of N forecasters, who share a common prior probability q that the state is good. This results in a prior belief that the event E will occur with probability $Pr(E|q) = qg + (1-q)b$.[5] Each forecaster k receives a signal σ_k from $S \equiv \{s_1, \ldots, s_m\} \cup \{s_\emptyset\}$ regarding the underlying state. Without loss

[4] We discuss issues that arise with more states in Appendix B.

[5] As can be seen here, there is a one-to-one correspondence between the prior over states and the prior over events. A similar one-to-one correspondence exists over posteriors. In this section, we will use the words prior and posterior to refer to beliefs over both states and events and will differentiate between them if there is potential ambiguity. In the algorithm and empirical sections, the prior and posterior always refer to beliefs over events.

of generality, signals are normalized so that $s_i := p(\omega_G|s_i)$, where $p(\omega_G|s_i)$ is forecaster k's posterior belief on the probability of the true state being ω_G when $\sigma_k = s_i$. The uninformative signal satisfies $s_\emptyset := q$. Let $p(s_i|\omega)$ denote the probability of a signal s_i in state ω, satisfying $\sum_{s_i \in S} p(s_i|\omega) = 1$ for each $\omega \in \{\omega_G, \omega_B\}$. The conditional distribution of signals is represented by a likelihood matrix $[Q_{\omega j}]_{2\times(m+1)}$. The first and second rows give the likelihoods of each signal in states ω_G and ω_B respectively. Thus, $Q_{Gi} = Q_{1i} \equiv p(s_i|\omega_G)$. We will assume there exists at least one signal $s_l \in \{s_1, \ldots, s_m\}$, where $Q_{Gi} \in (0, 1)$, which implies that at least one signal provides noisy information about the underlying state. Consistent with our naming convention of states, we also assume $E[\sigma_k|\omega_G] > s_\emptyset > E[\sigma_k|\omega_B]$, which implies that signals are informative and the expected posterior belief is higher in the good state than the bad state.

Given a signal s_i such that $p(s_i|\omega_G) + p(s_i|\omega_B) > 0$, the posterior belief that the state is ω_G is given by

$$p(\omega_G|s_i) = \frac{p(\omega_G)p(s_i|G)}{p(\omega_G)p(s_i|\omega_G) + p(\omega_B)p(s_i|\omega_B)} = s_i.$$

A forecaster with signal σ_k predicts that the event E will occur with probability $Pr(E|\sigma_k) = \sigma_k g + (1 - \sigma_k)b$.

Each forecaster k is asked to report i) a *prediction* P_k on the probability of event E and ii) a *meta-prediction* M_k on the average of others' predictions. Since E is associated with the state, a forecaster's probability prediction is dependent on the forecaster's signal. 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 of event E, where

$$P(\sigma_k) = \sigma_k\, g + (1 - \sigma_k)\, b. \tag{3}$$

Also let P_k be the forecast of the k_{th} forecaster and let $\bar{P} = \frac{1}{N}\sum_{k=1}^{N} P_k$ be the average prediction. Forecaster k's meta-prediction is given by $M_k = \mathbb{E}[\bar{P}|\sigma_k]$.

For a given outcome state ω, the expected average prediction is given by $\mathbb{E}[\bar{P}|\omega] = \sum_{s_i \in S} p(s_i|\omega)[gs_i + b(1 - s_i)]$. Let $M(\sigma_k)$ denote the meta-prediction function. This function can be written as

$$M(\sigma_k) = \sigma_k \mathbb{E}[\bar{P}|\omega_G] + (1 - \sigma_k)\mathbb{E}[\bar{P}|\omega_B]. \tag{4}$$

The signal densities $\{Q_{Gi}, Q_{Bi}\}$, prior q, and state-conditional event probabilities $\{g, b\}$ are common knowledge to the forecasters but unknown to the aggregator.

Figure 1 plots $P(\sigma_k)$ and $M(\sigma_k)$ in the space of predictions and signals. We note three properties that are the basis for our recalibration algorithm. First, both functions increase linearly in σ_k with the prediction line being more steep than the meta-prediction line. Since $\mathbb{E}[\bar{P}|\omega_B] = M(0) > b$ and $\mathbb{E}[\bar{P}|\omega_G] = M(1) < g$, the average prediction will be underconfident in our setting in both states.

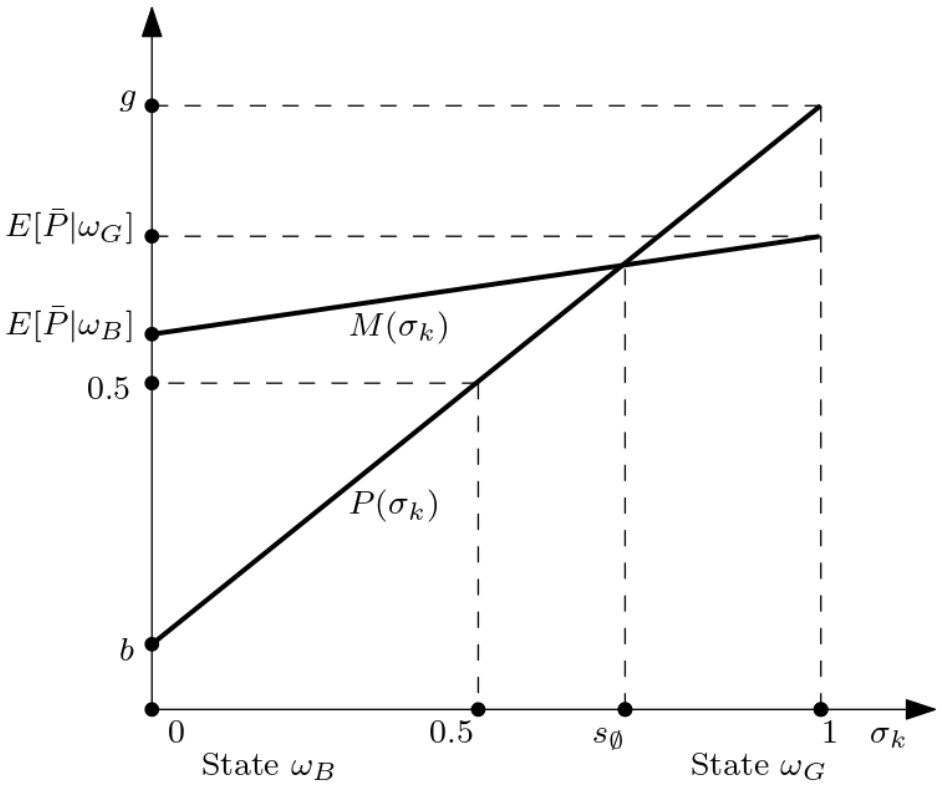


Figure 1: Example prediction and meta-prediction functions. Note that, in this example case, the average forecast is higher than 0.5 in both the good and the bad state. Section 3 will explore a potential pitfall in recalibrating such forecasts.

Second, the prediction and meta-prediction correspondences cross exactly once. The following lemma, proved in Appendix A, shows that this crossing point occurs at the uninformative prior.

Lemma 1. *$M(s_\emptyset) = P(s_\emptyset)$, i.e. a forecaster k's meta-prediction is equal to her prediction at the prior.*

As seen in the proof of Lemma 1 in Appendix A, the cross point property is due to the posteriors being a mean preserving spread of the prior. This is a very general characteristic of signals and the previous literature has highlighted how it leads to robustness in a variety of algorithms that elicit probabilities and meta-predictions. Wilkening et al. (2022) show that the crossing property holds in decision problems where probability forecasts are miscalibrated as long as miscalibrated forecasts are common knowledge. Chen et al. (2021) show that the crossing continues to hold in decision problems where signals are correlated.[6,7]

Finally, since both lines are linear, it is possible to identify $P(s_\emptyset)$ when there are at least two forecasters with different signals using the crossing point property and a projection. To see this, note that it is possible to rewrite the prediction function as:

$$\sigma_k = \frac{P(\sigma_k) - b}{g - b}.$$

Substituting this in Equation 4 yields

$$M(\sigma_k) = \alpha(Q, q, g, b) + \beta(Q, q, g, b)P(\sigma_k),$$

[6]Both of these papers explore prediction algorithms that try to correctly predict the correct state rather than make a probabilistic forecast. Wilkening et al. (2022) uses the ordering of the average prediction and average meta-prediction to the left and the right of the prior to make predictions. Chen et al. (2021) predicts $\mathbb{E}[\bar{P}|\omega]$ in each state using the relationship between predictions and meta predictions and selects the state where the average prediction is closest to the predicted average.

[7]When there are more than two states, a given signal might indicate a higher poster likelihood for multiple states. This can lead the prediction and meta-prediction lines to be non-linear. Thus, while the two lines continue to cross at the prior, the cross point may not be unique. In the appendix, we show example settings where there exists additional states that lead to a highly non-linear prediction and meta-prediction function.

where $\alpha(Q,q,g,b) := \frac{g\mathbb{E}[\bar{P}|\omega_B] - b\mathbb{E}[\bar{P}|\omega_G]}{g-b}$ and $\beta(Q,q,g,b) := \frac{\mathbb{E}[\bar{P}|\omega_G] - \mathbb{E}[\bar{P}|\omega_B]}{g-b}$ are constants that do not vary with σ_k. Using any two forecasts and meta-predictions that differ, the terms $\alpha(Q,q,g,b)$ and $\beta(Q,q,g,b)$ can be solved. $P(s_\emptyset)$ can then be identified by finding the point where $M(s_\emptyset) = P(s_\emptyset)$.

3 Robust recalibration

As discussed in Section 1, the traditional approach to extremizing uses the average probability of 0.5 as the threshold for determining whether forecasts are extremized towards 0 or 1. This approach can improve forecasts that are underconfident, but problems can arise in some settings where the prior is not 0.5.

Figure 1 illustrates the potential problem. The prior is biased towards true and the average prediction in the bad state is above 0.5. As seen in Equation 1, the LLO transformation leads to either $t(\bar{P}) > \bar{P} > 0.5$ or $t(\bar{P}) < \bar{P} < 0.5$ for $\bar{P} \neq 0.5$. Figure 1 depicts an example where $E[\bar{P}|\omega_B] > 0.5$ while $b < 0.5$. Thus, in state ω_B, $t(\bar{P})$ is expected to be even more inaccurate than the original average probability.

Definition (Wrong-sided average prediction)**.** *Average prediction $\bar{P}$ is wrong-sided if i) $\omega = \omega_G$ and $\bar{P} < 0.5 < g$ or, ii) $\omega = \omega_B$ and $\bar{P} > 0.5 > b$.*

Extremization away from 0.5 increases the miscalibration in a wrong-sided average prediction. When can the average prediction be wrong sided? First, it must be the case that $P(s_\emptyset) \neq 0.5$ for forecasts to be wrong-sided as the sample size grows large. This implies that the decision problem is *biased* in the sense that an infinite sample of uninformative forecasters would still predict one outcome. To see this, note that in a two-state environment, $E[\bar{P}|\omega_B] < P(s_\emptyset) < E[\bar{P}|\omega_G]$ and thus, in an unbiased problem, the average prediction will always coincide with the state as the number of forecasters grows large. Second, wrong-sidedness can only occur in one of the two states. This again follows from the fact that the prior is always between 0 and 1 and the posteriors are a mean preserving spread of the

prior. This implies that on average extremitization away from 0.5 can still be beneficial (as found in the literature) but suggests that an algorithm that better identifies cases where wrong-sidedness may occur can improve outcomes.

To account for situations where the average prediction can be wrong-sided, we propose the following **Robust Recalibration** procedure. First, using the data we estimate the meta-prediction function.

$$M_k = \beta_0 + \beta_1 P_k + \epsilon.$$

Denoting the estimates $\{\hat{\beta}_0, \hat{\beta}_1\}$, the predicted probability at the prior is found by finding the probability where the prediction and meta-prediction are equal. This will be given by $\hat{P}(s_\emptyset) = \hat{\beta}_0/(1-\hat{\beta}_1)$ for $\hat{\beta}_1 \neq 1$.

Next, using the estimated uninformed prediction $\hat{P}(s_\emptyset)$, we propose a transformation function $t_r(\bar{P})$ that satisfies the following function:

$$log\left(\frac{t_r(\bar{P})}{1-t(\bar{P})}\right) = log\left(\frac{\bar{P}}{1-\bar{P}}\right) + \gamma\left[log\left(\frac{\bar{P}}{1-\bar{P}}\right) - log\left(\frac{\hat{P}(s_\emptyset)}{1-\hat{P}(s_\emptyset)}\right)\right]. \tag{5}$$

Equation 5 suggests a linear transformation in log odds where (i) $\bar{P} \geq \hat{P}(s_\emptyset)$ is adjusted towards 1 and (ii) $\bar{P} < \hat{P}(s_\emptyset)$ is adjusted towards zero 0 when $\gamma \geq 1$. Note that for $\hat{P}(s_\emptyset) = 0.5$, Equation 5 is the same as Equation 2 with a reparametrization of the slope—$1+\gamma$ instead of γ—and an intercept of zero. Thus, in the special case of the estimated prior being unbiased ($\hat{P}(s_\emptyset) = 0.5$), t_r reduces to the LLO transformation away from 0.5 with $\delta = 1$, also known as the Karmarkar equation (Karmarkar, 1978).

Solving Equation 5 for $t_r(\bar{P})$, we get

$$t_r(\bar{P}) = \frac{\delta \bar{P}^{1+\gamma}}{\delta \bar{P}^{1+\gamma} + (1-\bar{P})^{1+\gamma}} \tag{6}$$

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

Section 4 tests the robust recalibration method $t_r(\bar{P})$ using a variety of experimental data sets. Note that the case of $\hat{P}(s_\emptyset) = 0.5$ (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 $\hat{P}(s_\emptyset) = 0.5$ 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 $P(s_\emptyset)$ improves the calibration. The analysis will compare t_r with various aggregation mechanisms that generate probability forecasts.

4 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 4.1 introduces the data sets. Section 4.2 presents preliminary evidence on the existence of wrong-sided average predictions and discusses estimated priors. Section 4.3 offers a comparative analysis on the calibration of transformed probabilities[8].

4.1 Data Sets

We investigate the empirical performance of robust recalibration using four distinct types of experimental tasks. Our data sources are Wilkening et al. (2022) and Howe et al. (2023). Participants in Wilkening et al. (2022) were presented with simple true/false scientific state-

[8]R software environment is used for empirical analysis (R Core Team, 2023; RStudio Team, 2020; Wickham, 2007; Wickham et al., 2019; Neuwirth, 2022).

ments. 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. Wilkening et al. (2022) collected data from 500 such statements. Howe et al. (2023) replicated the experiment using a subset of these statements. Each implementation recruited a new sample of subjects. Thus, we treat each statement-forecasting crowd combination as a distinct 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.

Howe et al. (2023) collected predictions and meta-predictions on various other 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 we use in our study 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. Sample size per task is 98 or 99. Appendix C provides a sample of tasks from each data set. We have a grand total of 910 tasks in our data.

4.2 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.

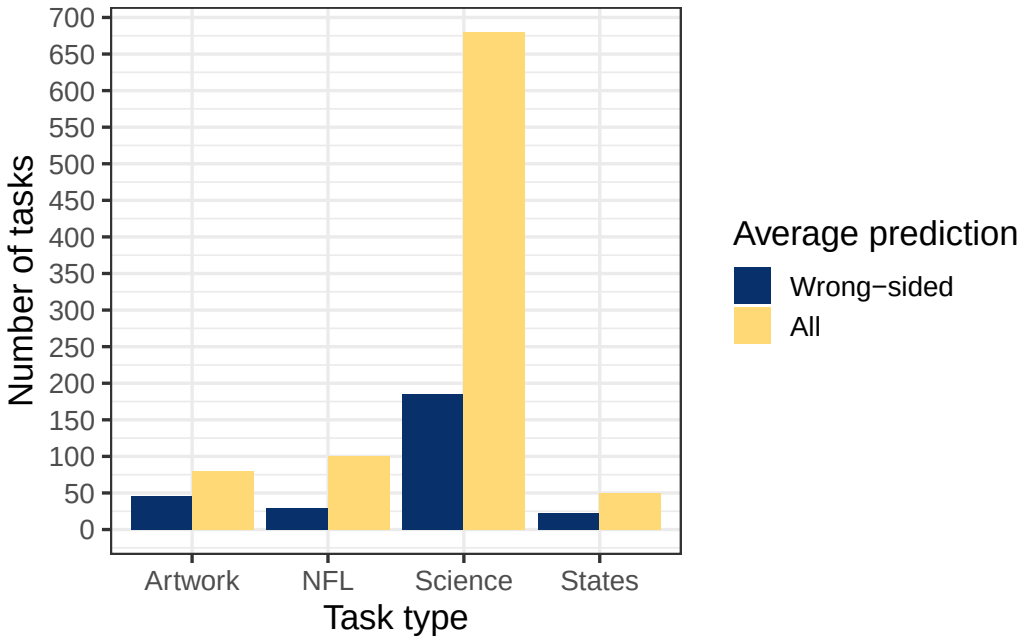


Figure 2: Wrong-sided averages in each data set

Figure 2 demonstrates that the average prediction is wrong-sided in a considerable number of tasks in each data set. In Figure D2 in Appendix D, we split the analysis into "True" and "False" statements. As seen there, 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".

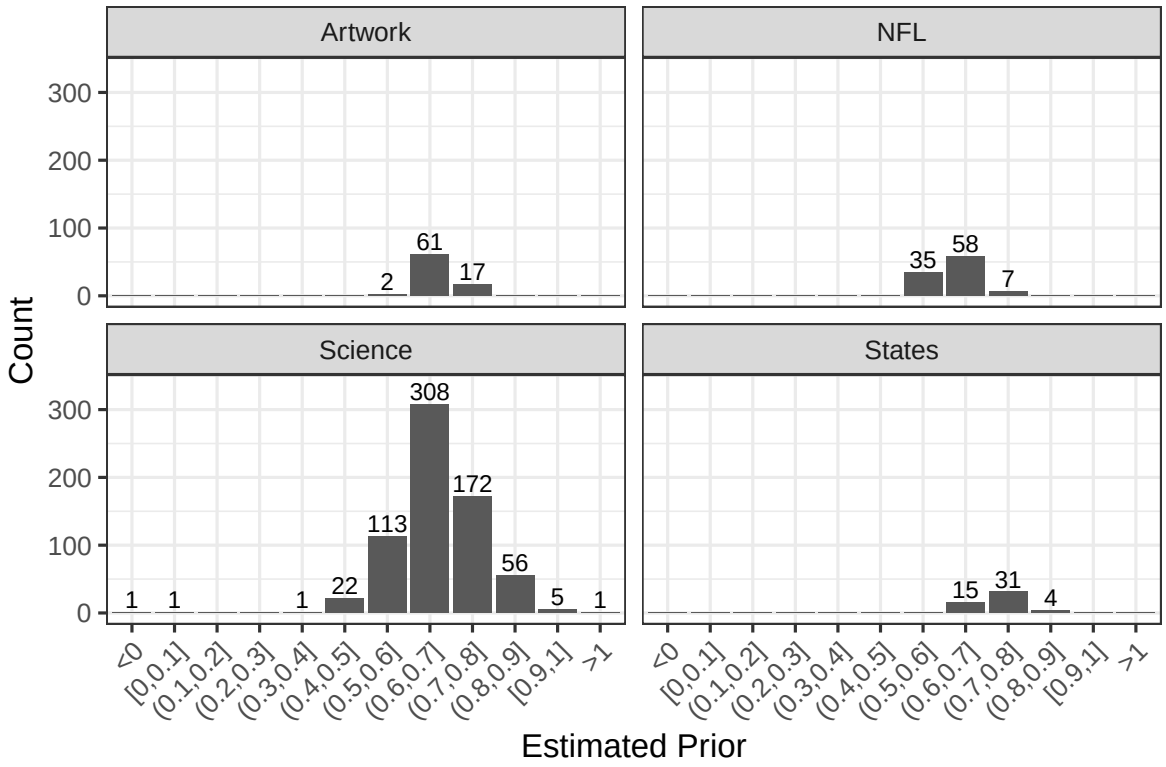


Figure 3: The distribution of estimated priors in each data set.

Figure 3 estimates the prior using the first stage of our robust recalibration procedure and supports the conjecture that there is a bias towards true. 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 $P(s_\emptyset)$ accurately. In our implementation, we set $\hat{P}(s_\emptyset) = 0.5$ in the two instances where $\hat{P}(s_\emptyset) \notin (0, 1)$. In other words, robust recalibration reverts to simple extremization when $\hat{P}(s_\emptyset) \notin (0, 1)$.

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

(a)

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

(b)

	Correct answer: True	Correct answer: False	Total
$\bar{P} > \hat{P}(s_\emptyset)$	270	40	310
$\bar{P} < \hat{P}(s_\emptyset)$	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. Thus, 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 4.3 implements extremization, robust recalibration and various aggregation algorithms in our data sets to investigate the calibration of the resulting probabilities.

4.3 Results

This section investigates the accuracy and calibration of the robust-recalibrated probability forecasts. We 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 6 with $\hat{P}(s_\emptyset) = 0.5$, which corresponds to $\delta = 1$. We consider $\gamma \in \{1, 1.5, 2, 2.5\}$ in our implemen-

tations of Equation 6 for both extremization and robust recalibration. Our second analysis compares robust recalibration to various 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.

4.3.1 A comparison of Robust Recalibration to the average prediction and the average extremized away from 0.5

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 4×3 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 "True", while for an average probability below 0.5, one minus the probability gives the confidence for "False". The coloring in Figure 4 breaks down the distribution of score for five different confidence levels of the corresponding average prediction. 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). The robust recalibration procedure offers additional improvements in Brier score over both the average and standard extremization approach for all potential γ parameters that we explored. As seen in Table 2, the performance difference between extremization and robust recalibration is significant in a paired Wilcoxon sign rank test that treats each decision problem as an observation.

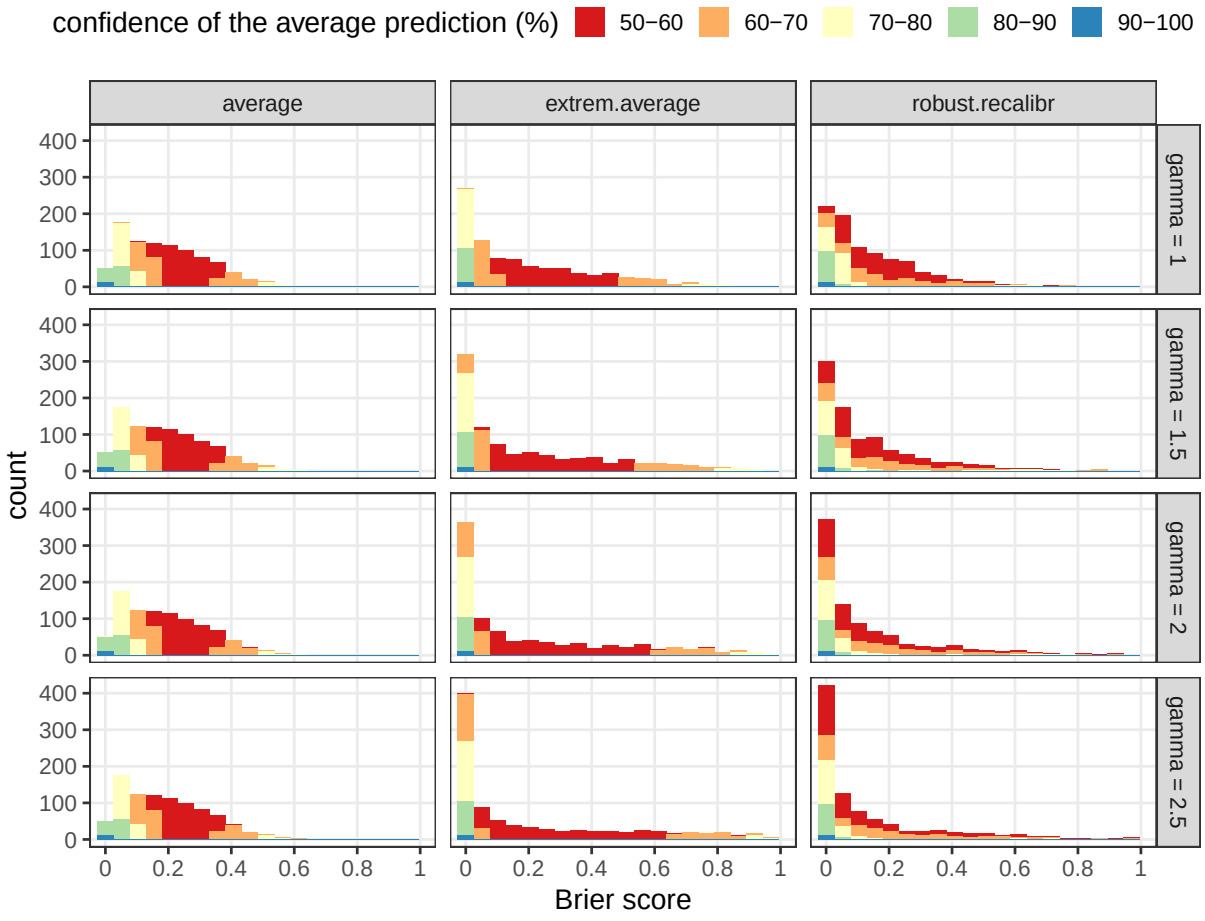


Figure 4: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

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	$V = 154699$	$p < 0.0001$

Table 2: Two-sided paired Wilcoxon signed rank test of Brier scores, Robust recalibration vs Extremizing away from 0.5.

Figure D1 in Appendix D extends the analysis here and graphs pairwise difference in Brier scores between extremization and robust recalibration with problems coloured by the average prediction. As seen there, robust recalibration is particularly effective in transforming low-confidence average predictions. 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 D2 in Appendix D 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.

We explored this issue by plotting calibration curves for the simple average, extremized average and robust recalibration. Calibration curves were constructed for each method by first separating the data into bins of $\{[0, 0.1], (0.1, 0.2], \ldots, (0.9, 1]\}$ based on the predictions of each method. We then plotted the predicted probability of "True" in each bin against the actual proportion of problems where true was the correct answer.

Figure 5 shows the calibration curves with a separate panel for each γ in the analysis set. 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.

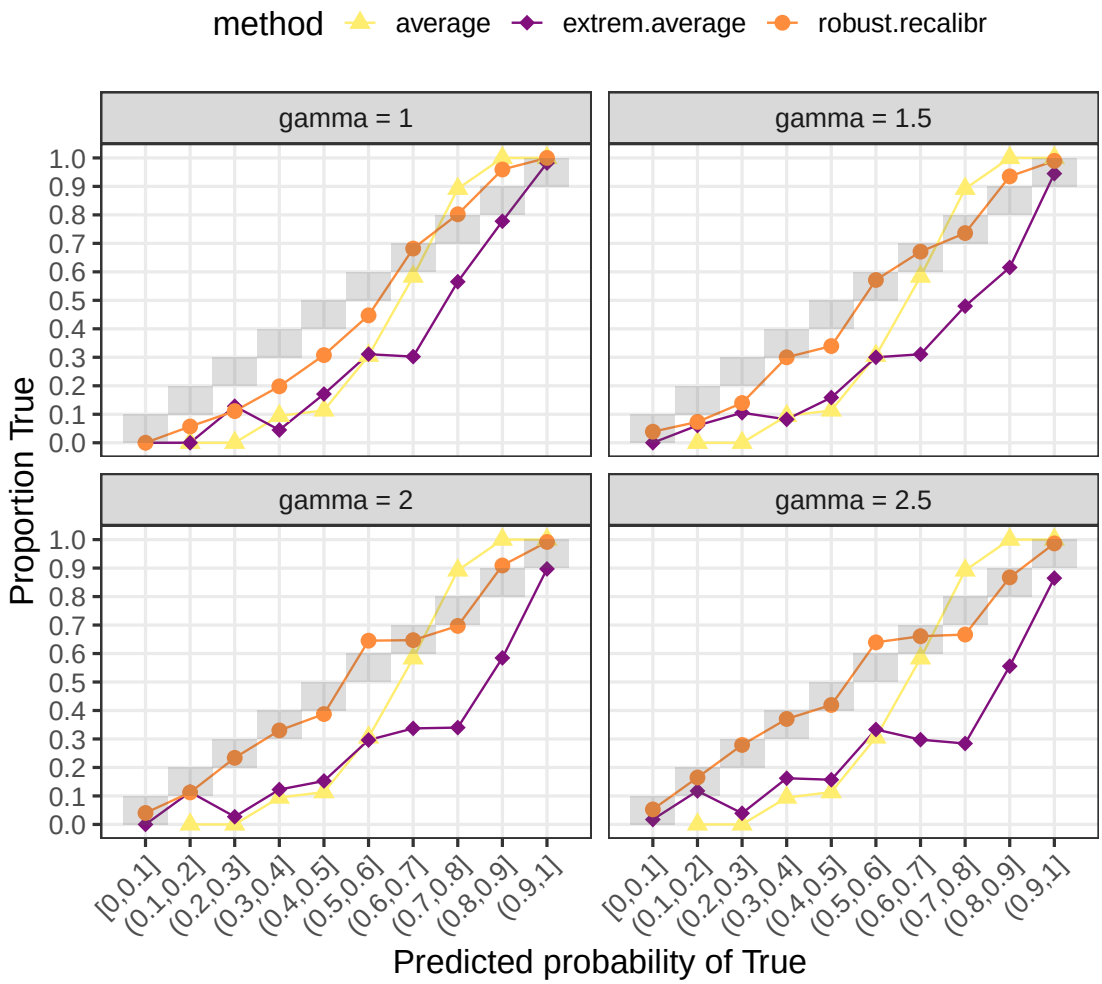


Figure 5: Calibration curves for simple average, extremized average and robust-recalibrated probabilities.

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.

4.3.2 A comparison of Robust Recalibration to other forecasting algorithms that use meta-predicitons

Our analysis thus far compared robust recalibration to methods that do not use meta-prediction data. One might wonder how it performs against alternative existing methods that seek to use meta-predictions to produce forecasts. To answer this question, we formed predictions using a number of alternative algorithms that exist in the literature. We elaborate on how these algorithms were constructed before continuing on to our second comparative analysis.

We consider four alternative algorithms that seek to exploit meta-predictions to improve forecasts:

1. **Meta-probability weighting:** This algorithm 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.

2. **Knowledge-weighting:** This algorithm, developed in (Palley & Satopää, 2023), seeks to construct optimal weights that minimize the "peer-prediction gap". This gap 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).

3. **Minimal pivoting:** This algorithm 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.

4. **Surprising Overshoot (SO) algorithm:** This algorithm is another aggregation method that addresses the shared-information problem (Peker, 2023). Information available to a forecaster determines the meta-prediction as well as the prediction, result-

ing 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 for each of the benchmark alorithms and our robust recalibration method with $\gamma = 2$. Similar to Figure 4, we color-coded the confidence levels of the average prediction in the corresponding prediction task to identify potential patterns over types of decision problems. Figure 6 demonstrates that 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 = 5.875, p = 0.000103). Pairwise comparisons suggest that robust recalibration achieves the highest accuracy [9].

In addition to the Brier score, we also constructed the calibration curve for each algorithm to understand how each algorithm is reshaping the predictions. These calibration curves are presented in Figure 7 and were constructed using the same methodology as Figure 5. As seen in the diagram, robust recalibration achieves the best calibration among alternatives for 9 out of 10 bins and is the second best in the remaining bin.

[9] 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, $p = 0.0075$; mean diff = -0.0223 vs surp.overshoot, $p = 0.0151$.

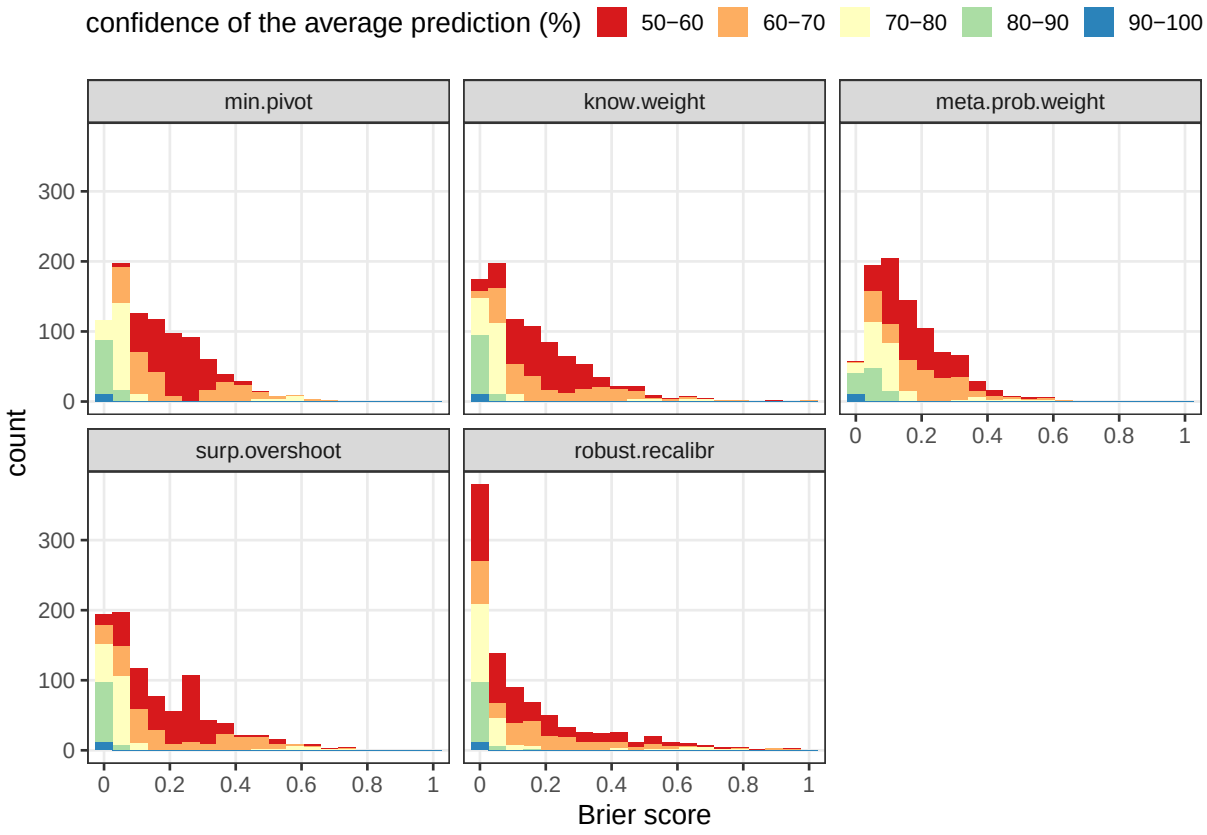


Figure 6: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

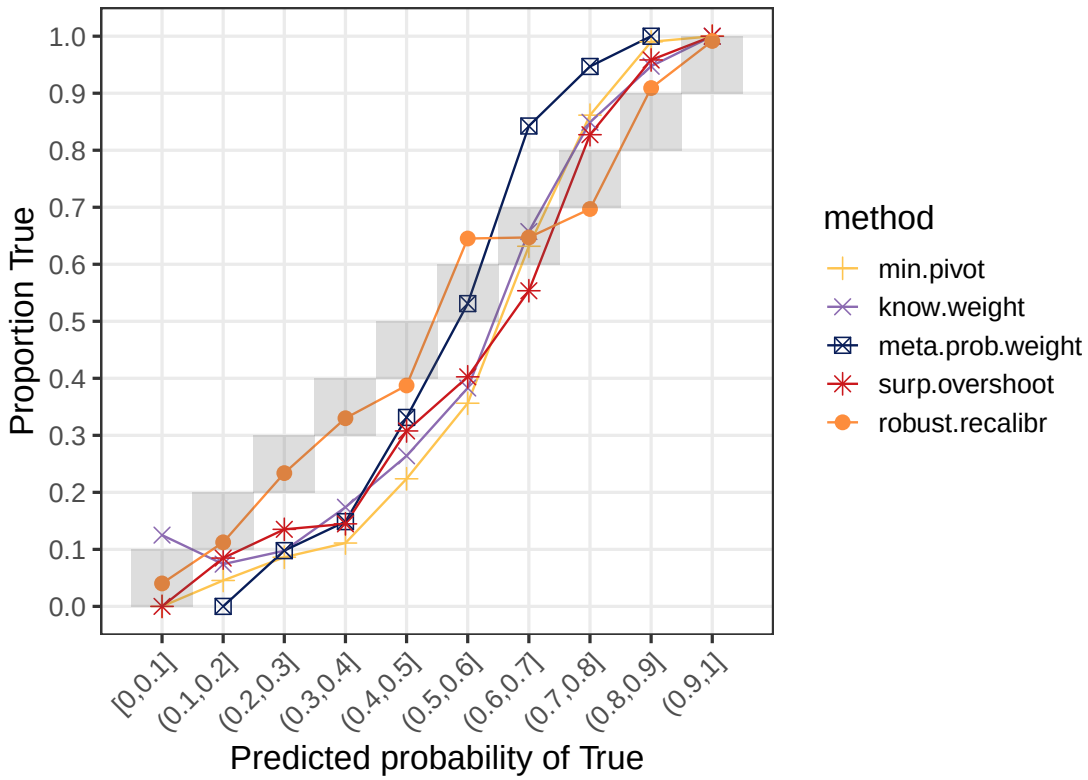


Figure 7: Calibration curves for simple average, extremized average and robust-recalibrated probabilities.

5 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.

Chen, Y.-C., Mueller-Frank, M., & Pai, M. M. (2021). *The wisdom of the crowd and higher-order beleifs.* mimeo.

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).

Libgober, J. (2023). *Identifying wisdom (of the crowd): A regression approach.* mimeo.

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.

Neuwirth, E. (2022). Rcolorbrewer: Colorbrewer palettes [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=RColorBrewer` (R package version 1.1-3)

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. (2023). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*, *94*(3), 467–501.

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

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

RStudio Team. (2020). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from `http://www.rstudio.com/`

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.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from `http://www.jstatsoft.org/v21/i12/`

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

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 Proofs

Proof of Lemma 1: This result is due to the fact that posteriors are a mean-preserving spread of the prior. At the prior:

$$\begin{aligned} P(s_\emptyset) = P(E|\sigma_k = s_\emptyset) &= \sum_i [P(E|s_i)P(s_i|s_\emptyset)] \\ &= \sum_i [qP(E|s_i)P(s_i|\omega_G) + (1-q)P(E|\sigma_i)P(s_i|\omega_B)] \\ &= q\sum_i [P(E|s_i)P(s_i|\omega_G)] + (1-q)\sum_i [P(E|s_i)P(s_i|\omega_B)] \\ &= q\mathbb{E}[\bar{P}|\omega_G] + (1-q)\mathbb{E}[\bar{P}|\omega_B]. \end{aligned}$$

Next, we note that $M(s_\emptyset) = E[\bar{P}|s_\emptyset]$. Thus, we can write

$$\begin{aligned} M(s_\emptyset) &= \mathbb{E}[\bar{P}|s_\emptyset] \\ &= \mathbb{E}[\bar{P}|\omega_G]p(\omega_G|s_\emptyset) + \mathbb{E}[\bar{P}|\omega_B]p(\omega_B|s_\emptyset) \\ &= q\mathbb{E}[\bar{P}|\omega_G] + (1-q)\mathbb{E}[\bar{P}|\omega_B]. \end{aligned}$$

Noting that the last expressions are the same, it follows that $P(s_\emptyset) = M(s_\emptyset)$. ■.

B Robust Recalibration with more than two states

In the main text, we showed that it is always possible to correctly estimate the prior using prediction and meta-predictions in an environment where there is exactly two states. This ensured that the algorithm would always identify the correct direction for extremitization in large sample. In this appendix, we use two examples to show that this the properties of the algorithm are not guaranteed when there are more than two states. The first example shows that the predition and meta-prediction lines may cross multiple times when we increase the state space and that the estimated prior may not be correct. Nonetheless, the algorithm may still function well as long as the estimated prior still identifies the correct direction for extremitization.

The second example identifies a situation where our algorithm fails to extremitize in the correct direction for one of the states. The counter-example highlights a case where the monotone likelihood ratio principal is violated and where signals are very informative about the signals of others but only weakly informative about the underlying likelihood of an event. In such cases, it is possible to construct situations where the meta-prediction line is non-linear and create perverse cases where the algorithm fails. We see such situations as being quite rare, but the possibility of such cases warrant an empirical exploration of the algorithm as is done in section 4 of the paper.

In both examples, we use a general likelihood matrix $\mathbf{Q}$ where the rows correspond to states and the columns relate to signals. Predictions and meta-predictions can be written using the posterior beleiefs for each state just as in the main text.

Example 1: Multiple Cross Points where the estimated posterior is incorrect but the direction of extremitization is correct. Suppose there are four states with probabilities of E given by $\{.8, .6, .4, .2\}$. For simplicity, we will refer to the states by using the corresponding probability. Forecasters have a prior of $\{1/4, 1/4, 1/4, 1/4\}$ over the states. Each forecaster receives a signal from $\{s_1, s_2, s_\emptyset, s_3, s_4\}$. The likelihood matrix is given by

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}.$$

Rows 1 to 4 (top to bottom) give the likelihoods for states $0.8, 0.6, 0.4$ and 0.2 respectively while columns 1 to 5 (left to right) represents the signals $s_1, s_2, s_\emptyset, s_3$ and s_4. Unlike the binary framework, the signals do not represent the posterior beliefs on one of the states. However, signals with a higher index indicate a weakly higher posterior probability on the "best" state (i.e. state 0.8). In this example, $\{s_3, s_4\}$ are generated when we are in state .8 or .6, while $\{s_1, s_2\}$ occur in states .4 and .2. Posterior belief on state 0.8 is highest for s_4, followed by s_3 and s_1, s_2 where the last two imply zero probability. Figure B1 depicts the corresponding prediction and meta-prediction functions.

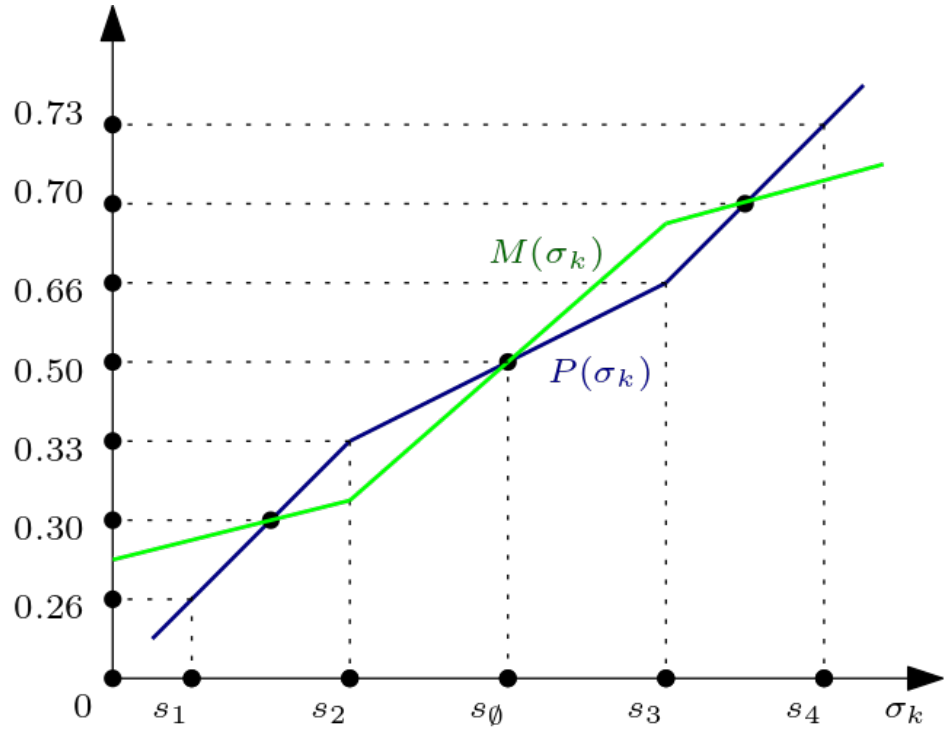


Figure B1: Example 1 prediction and meta-prediction functions (linear extrapolations from the predictions and meta-predictions at $\sigma_k \in \{s_1, s_2, s_\emptyset, s_3, s_4\}$).

The prediction and meta-prediction functions intersect at two distinct values other than $s_\emptyset$. Thus, solving for $M(x) = P(x)$ does not uniquely recover the prior. Nevertheless, this example demonstrates that robust recalibration could transform the average in the correct direction despite the inaccuracy in estimating $s_\emptyset$. To see this, we first calculate the average prediction, which are $\{0.71, 0.69, 0.31, 0.29\}$ in states $\{0.8, 0.6, 0.4, 0.2\}$ respectively. If the true state is 0.2 or 0.4, we get $\sigma_k \in \{s_1, s_2\}$. Then, the estimated prior will be 0.3, as it would be the unique intersection of the prediction and meta-prediction functions in the corresponding range. Robust recalibration transforms 0.29 and 0.31 away from 0.3, which could lead to transformed probabilities closer to the true probability (0.2 and 0.4 respectively). In contrast, extremizing away from 0.5 adjusts 0.31 in the wrong direction in state 0.4. A similar result holds in states 0.6 and 0.8. Then, the estimated prior will be 0.7. Average predictions of 0.69 and 0.71 are robust-recalibrated in the correct direction while extremizing away from 0.5 pushes 0.69 further away from the true probability of the event in state 0.6.

Note that the robust recalibration procedure is effective even though it does not produce an accurate estimate of the actual prior ($P(s_\emptyset)$) in any state. The likelihood matrix suggests that the forecasters' have a non-zero posterior probability for two states only. The prediction and meta-prediction functions are locally linear and estimated prior gives the intersection.

Example 2: Violation of MLRP. Consider an example with three states with probabilities $\{0.7, 0.4, 0\}$. Forecasters have a uniform prior $\{1/3, 1/3, 1/3\}$ over the states. Each forecaster receives a signal from $\{s_1, s_\emptyset, s_2, s_3\}$ according to the following likelihood matrix:

$$\mathbf{Q} = \begin{bmatrix} .3 & 0 & \frac{1}{3} & .367 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ .7 & 0 & 0 & .3 \end{bmatrix}$$

Rows 1 to 3 give the likelihoods of each signal in states 0.7, 0.4 and 0 respectively. Signals are ordered in the implied posterior belief on the best state (i.e. state 0.7) as $s_3 > s_2 > s_1$.

The prediction function satisfies $P(s_1) = 0.21$, $P(s_\emptyset) = 0.367$, $P(s_2) = 0.5$ and $P(s_3) = 0.39$. For meta-predictions, we first calculate the average prediction in each state, which leads to $E[\bar{P}|state = 0] = 0.264$, $E[\bar{P}|state = 0.4] = 0.463$ and $E[\bar{P}|state = 0.7] = 0.373$. For any agent with signal $\sigma_k \in \{s_1, s_\emptyset, s_2, s_3\}$, $M(\sigma_k)$ will be a convex combination of $E[\bar{P}|state]$ with weights being the posterior probabilities over the states. The resulting meta-prediction function satisfies $M(s_1) = 0.296$, $M(s_\emptyset) = 0.367$, $M(s_2) = 0.433$ and $M(s_3) = 0.37$. Figure B2 depicts the prediction and meta-prediction functions.

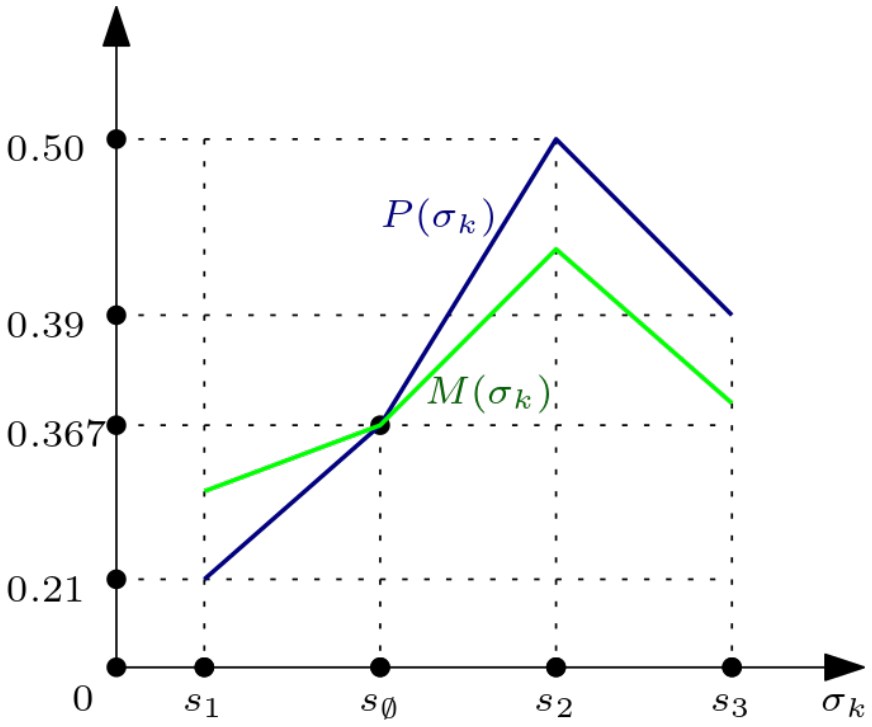


Figure B2: Example 2 prediction and meta-prediction functions

To see how robust recalibration performs, we randomly draw a sample of 10000 predictions and meta-predictions according to the functions in Figure B2. Then, we introduce random noise in meta-predictions and estimate the prior as described in Section 3. This procedure is repeated 100 times. Average estimated priors in each state is given by $\{0.366, 0.344, 0.357\}$ with standard errors strictly smaller than 0.001. Recall that the average predictions are 0.264, 0.463 and 0.373 in states 0, 0.4 and 0.7 respectively. Thus, the average should be recalibrated down in states 0 and 0.4 and up in state 0.7. Robust recalibration

transforms the average predictions in states 0 and 0.7 in the correct direction. However, in state 0.4, the robust recalibration procedure transforms the average in the wrong direction while extremization away from 0.5 would push the average towards 0.4.

The miscalibration in state 0.4 is a result of the intermediate signal being very informative about the predictions of others and the likelihood that the state is not 0. Recall that the posteriors in states $\{0.7, 0.4, 0\}$ following s_3 and s_2 are $\{0.367, 1/3, 0.3\}$ and $\{1/3, 2/3, 0\}$ respectively. Signal s_3 leads to the highest posterior on state 0.7 (followed by s_2 and s_1). However, s_2 rules out the worst state and leads to a higher probability prediction and meta-prediction overall. Since s_2 is more frequent in state 0.4, the resulting average prediction on the occurrence of the event is higher in state 0.4 than state 0.7, even though the event is more likely in the latter.

The miscalibration in this example would not occur if the likelihoods in state 0.4 are such that the resulting average prediction satisfies $E[\bar{P}] < E[\bar{P}|state = 0.4] < E[\bar{P}|state = 0.7]$. In the binary framework, signals can be normalized to represent the posterior beliefs on the good state (ω_G). Thus, higher expected signal in ω_G implies $E[\bar{P}|\omega_G] > E[\bar{P}|\omega_B]$. The same is not necessarily true for the "best state" in a multiple state framework where a signal is informative for beliefs on more than state. Note that the example considers a likelihood matrix where, given $s_3 > s_2 > s_1$, the expected signal is smaller in state 0.7 than state 0.4. In other words, the information in state 0.4 favors high states (and hence, a higher probability for the event) more than the information in state 0.7 on average. Such information structures are likely to be rare in practice, because it would imply that the evidence itself is expected to incorrectly suggest a higher probability in a lower state. Thus, we expect robust recalibration to perform well in most applications with more than two states.

C Prediction tasks

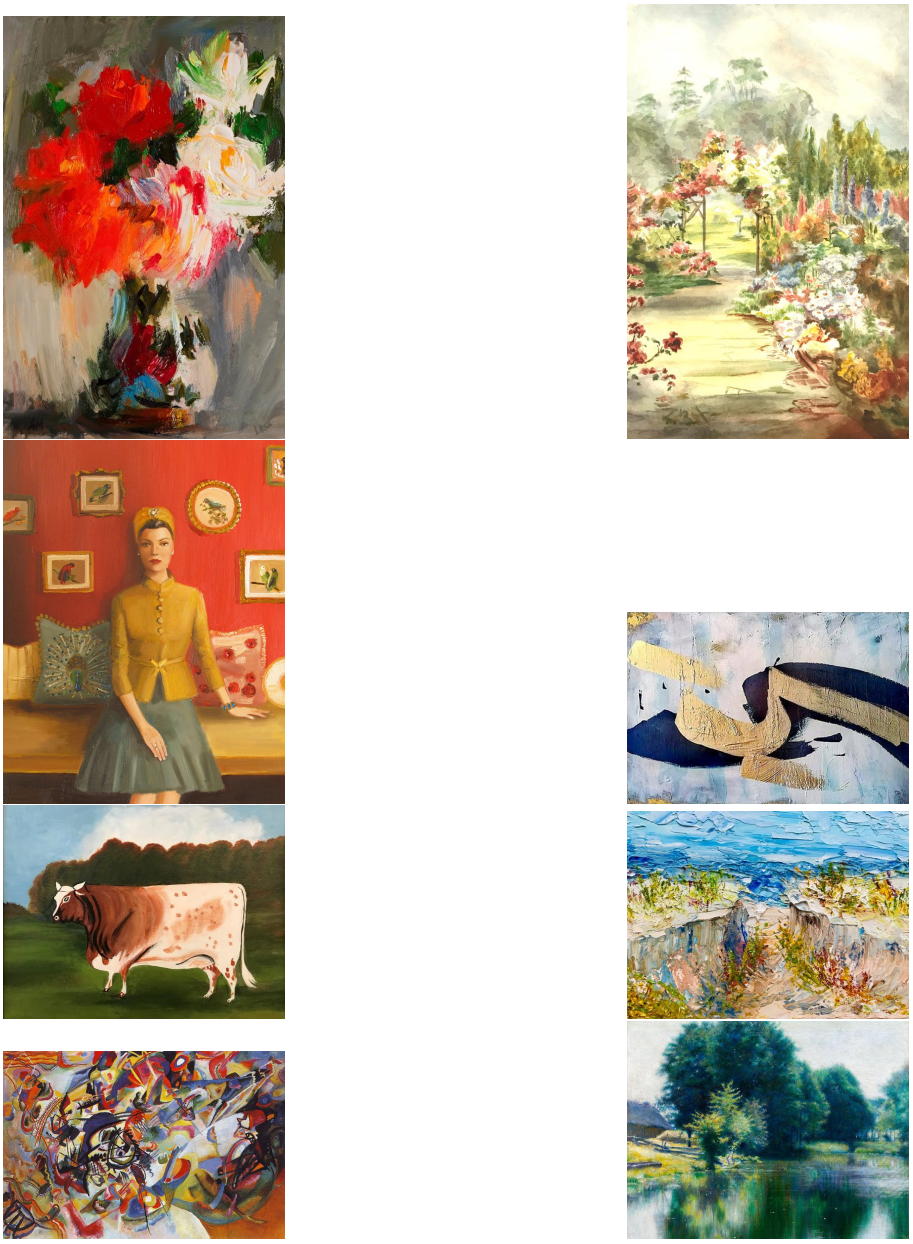
Table C1: Sample statements from Science and States data. See the supplemental material of Wilkening et al. (2022) for full list of statements

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 C2: 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 C1: Sample items from the Artwork data set



D Additional figures

Figure D1: 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. Negative differences indicate higher accuracy for robust recalibration.

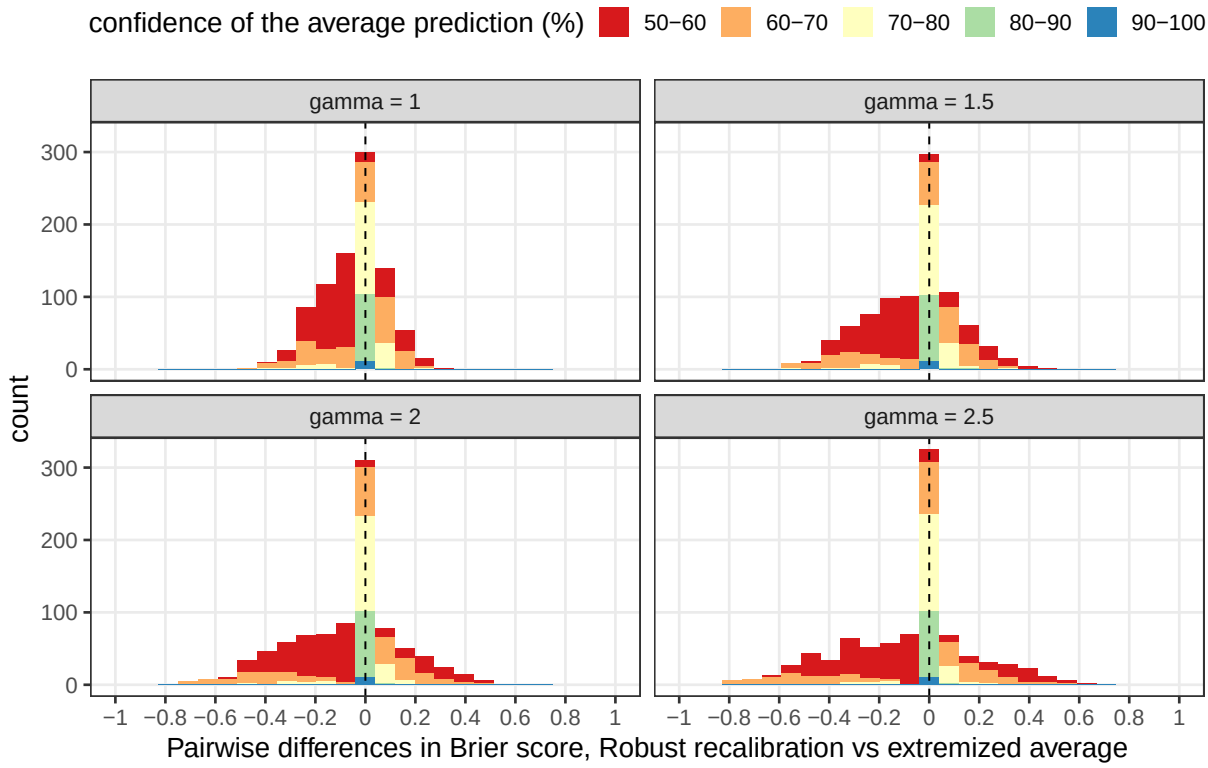


Figure D2: The distribution of average predictions for "True" and "False" statements in each data set.