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

August, 2024

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. Our algorithm works in single-question forecasting problems and does not require past data. 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) 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 often require probabilistic forecasts of uncertain events. Knowledge regarding the true likelihood of the event is often scattered across multiple individuals leading to an information aggregation problem where individual forecasts must be combined into a single forecast. Constructing the best aggregation method is difficult because forecasters may make errors when updating their information, may differ in expertise, and may vary in the overlap of the information they have available.

In data-rich environments, it is often possible to use information from training data or other forecasts to better understand the structure of information that exists amongst forecasters. In ideal settings, training data from past forecasts of known outcomes can be used to empirically estimate the diversity of information across individuals and aggregate unknown events accordingly (Atanasov et al., 2017; Breiman, 1996; Dana et al., 2019; Raftery et al., 1997; Satopää, Baron, et al., 2014; Satopää, Jensen, et al., 2014). Alternatively, in cases where forecasters are answering many questions, it may be possible to use answers from many questions to estimate features of the data-generating process that are useful to improving aggregation (Lichtendahl Jr et al., 2022; Satopää et al., 2017).

Unfortunately, decision-makers may not always have access to data that is relevant to the questions of importance. For example, the performance of forecasters on problems with known outcomes may not be relevant to the unknown problem of interest, and collecting relevant data on similar problems may be impractical (Clemen, 1989; Genre et al., 2013). The challenge in these "single-question" forecasting problems is to make the best forecast possible with data related only to the question being asked. We concentrate on the single-question problems for the rest of the paper.

The simple average is a common method to aggregate probability forecasts in the single-question domain (Winkler et al., 2019). 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, 2005). However, previous work suggests that the average probability forecast has a major shortcoming: 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). This aggregate conservatism naturally arises in settings where information is scattered and forecasters have access to different sets of information (Baron et al., 2014). It also arises even when individual forecasts are well-calibrated since the linear combination of probability forecasts is always theoretically miscalibrated and lacks sharpness (Ranjan & Gneiting, 2010).

One way to address the conservative bias is to recalibrate aggregate forecasts using an extremization function. 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 $\{\delta, \gamma\}$ are parameters.[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). However, a second potential issue arises in cases where the prior is biased. In many "wicked" forecasting problems, majority is wrong (Prelec et al., 2017; Wilkening et al., 2022) and/or inaccurate forecasters express higher confidence (Hertwig, 2012; Koriat, 2008, 2012; Lee & Lee, 2017). In these cases, 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.

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

In this paper, we ask whether it is possible to estimate the prior in a single-question framework and to use this as the starting point for recalibration. Our main contribution is to show that the common prior can be estimated in the single-question domain by eliciting forecasts and meta-predictions about the forecasts of others. We demonstrate how this information can be used to improve recalibration over standard single-question recalibration methods, and discuss its performance relative to other single-question algorithms that have recently been developed.

We consider an environment in which individuals share a common prior that an event may occur, which may be biased.[2] Forecasters receive independent signals conditional on the actual state, leading to an average probability forecast that 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, but the resulting forecast is always conservative. 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, even though the information contained in forecaster's private signals suggest a lower probability than the prior.

To address this issue, we elicit each forecaster's estimate on the average forecast of others (referred to as their meta-prediction) as well as their probabilistic forecast. We show that these two measures can be combined 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 robust recalibration algorithm calibrates, we estimate calibration curves across a variety of decision problems related to general knowledge, sports, and

[2]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 price of art works. For recalibration parameters in the range of those suggested in Baron et al. (2014), we find that our algorithm generally 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). Robust recalibration also generates very low brier scores across decision problems, suggesting that it has very good accuracy characteristics overall.

The rest of this paper is organized as follows: Section 2 reviews the recalibration literature and summarizes the other single-question algorithms that we compare our algorithm with. Section 3 introduces the Bayesian framework. Sections 4 discusses the existence of wrong-side average forecasts in biased decision problems and develops the robust recalibration method that utilizes meta-predictions. Section 5 provides empirical evidence from experimental prediction tasks. Section 6 provides an overview of our contribution and concludes.

2 Related Literature

Recalibration approaches that seek to account for the partial overlap in shared information amongst forecasters have been shown in a variety of settings to improve outcomes over techniques that allow only for a weighted average of individual predictions (Baron et al., 2014; Turner et al., 2014). Recalibration typically involves the use of an extremization function, which adjusts forecasts toward extreme outcomes. The most popular choices are logit and probit transformations (Baron et al., 2014; Satopää, Baron, et al., 2014; Satopää et al., 2016; Turner et al., 2014).

Recalibration functions are typically symmetric around 0.5. However, as noted in Turner et al. (2014), it is possible and often beneficial to allow for more flexible calibration approaches by extremizing from a different initial prior. A challenge in improving calibration is therefore to incorporate information about the prior into the aggregation algorithm (Dietrich, 2010; Satopää, 2022). This has been accomplished in multiple-question forecasting environments by using a Bayesian framework and multiple predictions within the same survey to estimate a non-uniform prior across a range of prediction tasks (Lichtendahl Jr et al., 2022; Satopää et al., 2017).[3]

Our approach within the recalibration literature is similar to Lichtendahl Jr et al. (2022), which also stress the importance of using a value other than 0.5 as the basis for extremization. In their paper, the authors explore data-generating processes in which experts observe multiple independent and identically distributed signals from a joint distribution along with multiple commonly observed private signals. The authors show that with multiple forecasts and historical data, it is possible to develop estimation procedures that are well calibrated and which "antiextremizes" the average in a large number of cases.

We see the empirical approach taken in Lichtendahl Jr et al. (2022) as being highly useful in environments where there is substantial historical data to estimate base rates and some confidence in the error structures generated from the data generating process. Our approach, which estimates the prior from meta-predictions and predictions alone, is likely to be more valuable in environments where forecasters have limited historical data and where there is significant uncertainly about the underlying data generating process. We note the two approaches are not mutually exclusive: it is an open and interesting question of how to best combine the two approaches when historical data, training data, and meta-prediction data are available.

Our paper also contributes to the emerging literature on forecast aggregation methods that rely on higher order beliefs (Chen et al., 2021; Martinie et al., 2020; Palley & Satopää, 2023; Palley & Soll, 2019; Peker, 2023; Prelec et al., 2017; Wilkening et al., 2022). The elicitation of higher-order beliefs allows the researcher additional information about the

[3]In settings where forecasters have heterogeneous preferences over the extent to which their forecast conforms or contrasts to the reports of others, it may also possible to estimate the prior using only choice data. See Jia et al. (2024) for an approach to improving forecasts in this alternative setting.

signals that individuals receive. Such information 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-prediction algorithms have been developed both for binary classification problems (e.g., Chen et al., 2021; Prelec et al., 2017; Wilkening et al., 2022) 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 the surprising overshoot (SO) algorithm. 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 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 individuals have only imperfect information about the true state. Palley and Soll (2019), Palley and Satopää (2023) and Peker (2023) use an alternative framework that allows 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.

Although it is not emphasized in the previous literature, the framework used in Palley and Soll (2019) is one in which the meta-prediction and prediction correspondences are linear and where the intersection of these lines corresponds to the common prior that exists after accounting for publicly observable information. As a result, the ordering of the prediction and meta-prediction correspondences switch at the uninformative prior. An implication of this is that the minimum pivoting mechanism—which uses the difference in the average prediction and meta-prediction to adjust forecasts—is fundamentally an extremizing procedure that adjusts forecasts away from the common prior. As seen in the results section, our algorithm with the suggested extremizing parameters of Baron et al. (2014) is more aggressive than the adjustment made in the pivot mechanism and performs better. Thus, at least in the data sets considered, our results suggest that the minimum pivot mechanism is too conservative. This finding is similar to the contemporaneous work presented in Rilling (2024) that explores a neutral pivoting mechanism that is more aggressive than the original minimal pivot mechanism.

Our recalibration procedure relies on a regression approach that is similar to the fitting technique used in Palley and 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.

3 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.

We consider a setting where there are two potential underlying states. Let $\omega \in \{\omega_G, \omega_B\}$ be the state of the world where G and B represent "Good" and "Bad" states respectively. Event E 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. Forecasters share a common prior that the state is good, q, resulting in a common prior belief that the event E will occur with probability $Pr(E|q) = qg + (1-q)b$.[4] 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 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 $s_\emptyset := q$ and the signal space is bounded in $[0,1]$.

Let $p(s_i|\omega)$ denote the probability of a signal s_i in state ω, satisfying $\sum\limits_{s_i \in S} p(s_i|\omega) = 1$ 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, $Q_{\omega_G i} = Q_{1i} \equiv p(s_i|\omega_G)$. We will assume there exists at least one signal $s_l \in \{s_1, \ldots, s_m\}$, where $Q_{\omega i} \in (0,1)$, which implies that at least one signal provides noisy information about the underlying state.[5] 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.

It is useful at this point to note a distinction that we are making regarding events and states. In our framework, the values b and g represent the best prediction that could be made by an aggregator in the corresponding state if he knew the structure of the information service

[4]As can be seen here, there is a one-to-one correspondence between the prior q on ω_G and the prior $qg + (1-q)b$ on the event E. A similar one-to-one correspondence exists between posteriors on ω_G and E. 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.

[5]This assumption implies that the signal distribution is non-degenerate in either state since $\sum_j Q_{\omega j} = 1$.

and observed an infinite number of draws from it. In some settings, such as asking about the answer to an objective true/false knowledge question, signals may be fully revealing and we could set g and b to 1 and 0 respectively. However, in other settings, such as predicting whether someone will be convicted of a crime, some aspects of the problem (e.g., the detailed knowledge of the jurists) may be unobservable. In these cases g and b represent the best possible predictions that could be made about the event based on all possible information available.

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

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

Given $p(\omega_G|\sigma_k) = \sigma_k$ for a forecaster with signal σ_k, posterior belief on the occurrence of event E is given by $Pr(E|\sigma_k) = \sigma_k g + (1 - \sigma_k)b$.

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. 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 the likelihood of E depends on 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

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

Further, let P_i be the prediction of forecaster i and let $\bar{P}_{-k} = \frac{1}{N-1} \sum\limits_{i \neq k} P_i$ be the average prediction made by the other $N - 1$ forecasters. Forecaster k's meta-prediction is given by $M_k = \mathbb{E}[\bar{P}_{-k}|\sigma_k]$.

For a given outcome state ω, the expected prediction made by a randomly selected other forecaster is given by

$$\mathbb{E}[P|\omega] \equiv \sum_{s_i \in S} p(s_i|\omega)[gs_i + b(1-s_i)].$$

Noting that we have assumed that signals are independent once we have conditioned on the state, $\mathbb{E}[\bar{P}_{-k}|\omega] = \mathbb{E}[P|\omega]$ for all k. Thus, the meta-prediction function, denoted by $M(\sigma_k)$, can be written as

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

Figure 1 plots $P(\sigma_k)$ and $M(\sigma_k)$ in the space of predictions and signals. We note three general 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. Note that $P(\sigma_k) \in [b, g]$ and $M(\sigma_k) \in [\mathbb{E}[P|\omega_B], \mathbb{E}[P|\omega_G]]$. We also have $\mathbb{E}[P|\omega_B] > b$ and $\mathbb{E}[P|\omega_G] < g$, i.e. the average prediction will be too conservative in our setting in both states. To illustrate, consider the case $\omega = \omega_G$ where the true probability of the event is g. Then, a forecaster k has a perfectly calibrated prediction $P(\sigma_k) = g$ only if $\sigma_k = 1$ and predictions are conservative for all $\sigma_k < 1$. Recall that at least one noisy signal about the state occurs with strictly positive probability by assumption. Thus, in a large enough sample, there will always exist forecasters with $\sigma_k < 1$, leading to an average prediction lower than g. Furthermore, it is common knowledge that forecasters with $\sigma_k < 1$ exist. Forecasters with $\sigma_k = 1$ expect average prediction to be more conservative than their own prediction, implying $M(\sigma_k) < P(\sigma_k) = g$ for $\sigma_k = 1$. A similar reasoning holds for $\omega = \omega_B$, resulting in conservatism in average prediction and a relatively more steep prediction line.

Second, the prediction and meta-prediction lines cross exactly once. Figure 1 illustrates this result. Both functions are monotonically increasing, linear in σ_k, and the range of meta-predictions is a subset of predictions, resulting in a unique crossing point. Lemma 1 shows that this crossing point occurs at the uninformative prior. All proofs are included in Appendix A.

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

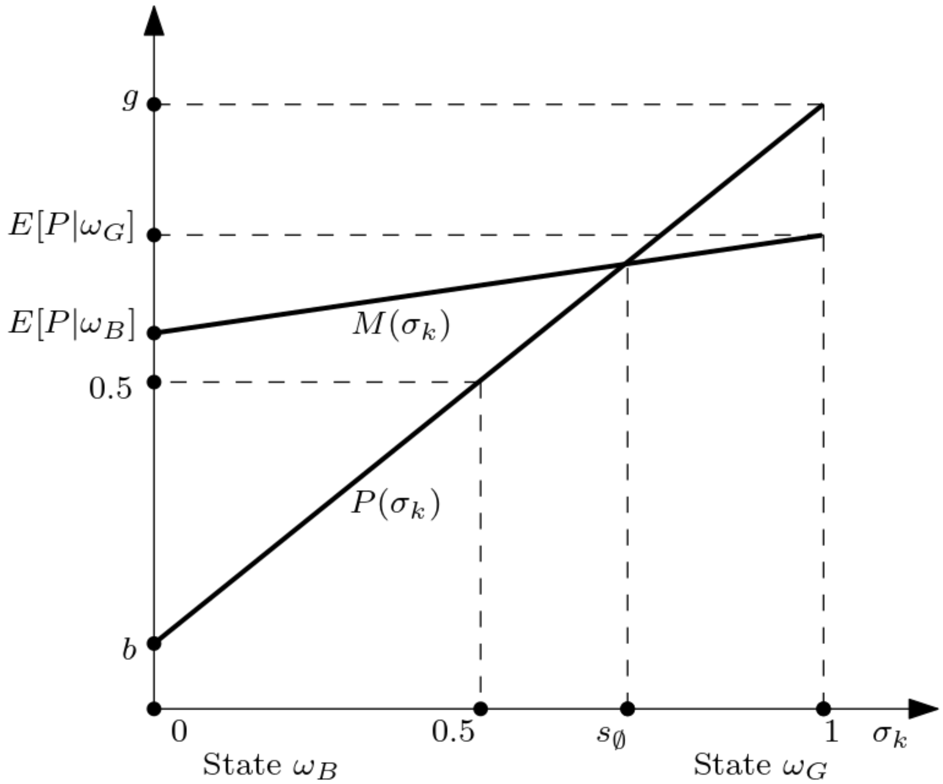


Figure 1: Prediction and meta-prediction functions for a case of $s_\emptyset > 0.5$. Note that, in this example, the average forecast is higher than 0.5 in both the good and the bad state. Section 4 will explore a potential pitfall in recalibrating such forecasts.

at the prior.

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

$$M(\sigma_k) = \alpha(Q, q, g, b) + \beta(Q, q, g, b)P(\sigma_k), \tag{5}$$

where $\alpha(Q, q, g, b) := \dfrac{g\mathbb{E}[P|\omega_B] - b\mathbb{E}[P|\omega_G]}{g - b}$ and $\beta(Q, q, g, b) := \dfrac{\mathbb{E}[P|\omega_G] - \mathbb{E}[P|\omega_B]}{g - b}$ 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. Prior prediction $P(s_\emptyset)$ can then be identified by finding the point where $M(s_\emptyset) = P(s_\emptyset)$.

Before turning to our recalibration strategy, we note that our model presents an ideal environment in which all forecasters perfectly map their signals to predictions and meta-predictions and there are exactly two states. Previous work suggests that the crossing point property between the meta-prediction and prediction correspondence is reasonably robust to systematic individual-level miscalibrations. 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 same property continues to hold in decision problems where signals are correlated.[6] Nonetheless, it is likely that there is idiosyncratic noise, particularly in the report of meta-predictions. As seen below, we use regression approaches to estimate the prediction and meta-prediction correspondences in order to help reduce the impact of such noise.

In Appendix B, we extend the theoretical discussion and provide two examples that show that the properties of the algorithm are not guaranteed when there are more than two states. The first example shows that the prediction and meta-prediction lines may cross multiple times when we expand the state space and that the estimated prior may not be correct. Nonetheless, the example demonstrates that the algorithm may still function well as long as the estimated prior still identifies the correct direction for extremization.

The second example identifies a situation where our algorithm fails to extremize in the correct direction for one of the states. The counter-example highlights a case where signals are very informative about the signals of others but only weakly informative about the underlying likelihood of the event. We see such situations as being quite rare: it requires

[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) use 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) predict $\mathbb{E}[\bar{P}|\omega]$ 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.

a very specific signal structure where the event of interest is only weakly connected to the signals. Nonetheless, the possibility of such cases warrants a careful empirical exploration of the algorithm to assess its applicability in real-world settings.

3.1 A special case with $b = 0$ and $g = 1$

In the empirical section below, we study true-false questions where there is an objectively correct answer. In these questions, it is possible that a very well-informed forecaster could know the state with certainty. Thus, these types of questions might be seen as a special case of our model where $b = 0$ and $g = 1$. In this special case, the prediction correspondence is $P(\sigma_k) = \sigma_k$, and the meta-prediction correspondence is as given by Equation 4 where $\mathbb{E}[P|\omega] \equiv \sum_{s_i \in S} p(s_i|\omega)s_i$ for $\omega \in \{\omega_G, \omega_B\}$. The prediction line is predicted to travel along the 45 degree line in the space of signals and predictions. Thus, the prior corresponds to the point where the meta-prediction correspondence crosses the 45-degree line.

In our empirical analyses, we do not directly impose that the prediction line is equal to the 45-degree line since testing this relationship would require information related to signals that are unobservable in empirical data. Instead, we estimate the two parameters in Equation 5 using linear regressions. We then use these estimates to predict the point where $M(s_\emptyset) = P(s_\emptyset)$. This approach is valid for any $0 \leq b < g \leq 1$ and therefore nests the special case where $b = 0$ and $g = 1$.

The second step of our algorithm involves extremizing the data away from this estimated prior. As discussed below, our algorithm can overshoot the true state when $0 < b < g < 1$ but not when $b = 0$ and $g = 1$. We therefore discuss the theoretical properties of the algorithm both for the general case and the special case below.

4 Robust recalibration

As discussed in Section 1, the traditional approach to extremizing compares the average probability prediction $\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$ to the threshold of 0.5 for determining whether forecasts are extremized towards 0 or 1. This approach can improve forecasts that are too conservative, 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 1 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 $E[P|\omega_B] > 0.5$ while $b < 0.5$. Thus, in state ω_B, $t(\bar{P})$ is expected to be even more inaccurate than the original average probability. We refer to such problems as being wrong sided:

Definition 1 (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 infinitely large. To see this, note that in a two-state environment, $E[P|\omega_B] < P(s_\emptyset) < E[P|\omega_G]$ and the average prediction will be the expected prediction in each state as the sample grows large. Second, wrong-sidedness can only occur in one of the two states. This follows from the fact that the prior is always between 0 and 1 and the expected posterior is equal to the prior. This implies that on average extremization 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 accuracy.

To account for situations where the average prediction can be wrong-sided, we propose the following **Robust Recalibration** procedure. We first use the data to estimate the prior. Following a similar approach to Palley and Satopää (2023), we allow for random noise ϵ in reported meta-predictions and assume:

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

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 $t_{RR}(\bar{P})$ that satisfies the following expression:

$$log\left(\frac{t_{RR}(\bar{P})}{1-t_{RR}(\bar{P})}\right) = log\left(\frac{\bar{P}}{1-\bar{P}}\right) + \gamma\left[log\left(\frac{\bar{P}}{1-\bar{P}}\right) - log\left(\frac{\hat{P}(s_\emptyset)}{1-\hat{P}(s_\emptyset)}\right)\right]. \tag{7}$$

Equation 7 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 $\gamma \geq 0$. Note that for $\hat{P}(s_\emptyset) = 0.5$, Equation 7 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_{RR} reduces to the LLO transformation away from 0.5 with $\delta = 1$, also known as the Karmarkar equation (Karmarkar, 1978).

Solving Equation 7 for $t_{RR}(\bar{P})$, we get

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

where $\delta = [(1-\hat{P}(s_\emptyset)/\hat{P}(s_\emptyset)]^\gamma$. Unlike simple extremization away from 0.5, $t_{RR}(\bar{P})$ 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_{RR} to adjust wrong-sided average predictions in the correct direction.

Our algorithm essentially uses two pieces of information to transform the average prediction. The first is the estimated common prior which reflects all the commonly shared information in the system. We treat this information as being important to prediction, but do not recalibrate it as it reflects information that is common across all forecasters. The second is the difference between the actual prediction and the common prior. This value reflects the average change in prediction based on the private signals available to the forecasters. As these signals are likely to have less overlap, using the average is likely to be conservative. Thus, by extremizing the difference, we hope to improve the outcome of the estimate.

We note an important factor in the relative performance of robust recalibration. The extent of transformation in robust recalibration depends on both γ and $\hat{P}(s_\emptyset)$, while simple extremization always uses $\hat{P}(s_\emptyset) = 0.5$. Thus, the step size for adjustment in the two methods may differ for the same value of γ. Note that for $0 < b < g < 1$, either method may over-adjust and produce more extreme probabilities than b and g in the corresponding state. The remainder of this section provides a comparative discussion on the properties of robust recalibration.

In problems that are wrong-sided, simple extremization will adjust predictions away from the true probability of the event while robust recalibration will adjust predictions in the direction of the true probability. As mentioned above, the extent of transformation is also a factor in accuracy. Proposition 1 compares simple extremization and robust recalibration in wrong-sided problems.

Proposition 1. *Suppose that the decision problem is wrong-sided. Then, there exists a threshold $g'(b')$ in state $\omega_G(\omega_B)$ such that, for all $g > g'$ $(b < b')$, robust recalibration leads to a lower average Brier score than extremization away from 0.5 for any identical tuning parameter γ in the limit as the sample size goes to infinity. The threshold becomes more extreme (g' to 1, b' to 0) as $|\bar{P} - P(s_\emptyset)|$ increases.*

Proposition 1 establishes that robust recalibration achieves higher accuracy in wrong-sided problems where the true probabilities in the good and bad state are sufficiently extreme. In other problems, however, there is the potential that robust recalibration "overshoots" the true probability. To illustrate with a numerical example, suppose $\omega = \omega_G$, $g = 0.55$,

$P(s_\emptyset) = 0.3$, $\bar{P} = 0.49$ and let $\gamma = 1$. Robust-recalibrated probability is 0.68, while simple extremization leads to 0.48. Robust recalibration transforms in the correct direction, but the overadjustment produces a less accurate prediction. Such overadjustment becomes more likely when the adjustment of robust recalibration is larger, which occurs when $\bar{P}$ is further away from $P(s_\emptyset)$.

The benefits of a proper probability transformation are highest when the true probability is close to 0 or 1. These problems represent situations where extremizing in the wrong direction is very costly in terms of accuracy and where there is little chance that robust recalibration overshoots the true probability. A special case where robust recalibration always improves the Brier score is one where $b = 0$ and $g = 1$. In these problems, it is not possible to overshoot the true probability through recalibration and a more extreme forecast is better on average as the sample grows large.

Proposition 2. *Suppose that the decision problem is wrong-sided, $b = 0$ and $g = 1$. Then robust recalibration leads to a strictly lower average Brier score than extremization away from 0.5 for any identical tuning parameter γ in the limit as the sample size goes to infinity.*

Proposition 2 follows from the observation that, unlike simple extremization, robust recalibration transforms wrong-sided average forecasts towards the correct extreme. Since over-adjustment is not a concern for $b = 0$ and $g = 1$, robust recalibration achieves strictly higher accuracy.

In decision problems where average forecast is not wrong-sided, both robust recalibration and simple extremization will adjust forecasts in the direction of the true state and therefore will lead to relatively similar forecasts. However, the intensity of adjustment could differ due to prior. This may affect the relative accuracy of the two algorithms depending on the extent to which the average forecast needs to be extremized. To illustrate, consider a simple example where $\omega = \omega_G$, $g = 0.75$, $\bar{P} = 0.6$, $P(s_\emptyset) = 0.4$ and $\gamma = 1$. As the sample of forecasters grow to infinity, robust recalibration recovers $P(s_\emptyset)$ and transforms according to Equation 8 with $\delta = 1.5$, which leads to a robust-recalibrated probability of 0.77. Simple extremization applies the same transformation with $\delta = 1$ and produces an extremized probability of 0.69. Since $g = 0.75$, robust recalibration achieves higher accuracy. Now suppose $P(s_\emptyset) = 0.55$ instead. Then, robust-recalibrated probability becomes 0.65 and simple extremization is more accurate. The opposite result would be true if g is closer to 0.5 and thus, requires a smaller extremizing adjustment. As a result, we can establish a general result only for the special case of $b = 0$ and $g = 1$.

Proposition 3. *Suppose that the decision problem is not wrong-sided, $b = 0$ and $g = 1$. Then, robust recalibration achieves a lower average Brier score than extremizing away from* 0.5 *for any identical tuning parameter γ if $|\bar{P} - P(s_\emptyset)| > |\bar{P} - 0.5|$ in the limit as the sample size goes to infinity.*

In Proposition 3, $|\bar{P} - P(s_\emptyset)| > |\bar{P} - 0.5|$ is simply a condition for a larger extremizing adjustment in robust recalibration than simple extremization. Since, extremizing is always beneficial and over-adjustment is not a concern, the algorithm with a more intensive extremization achieves higher accuracy.

Taking these propositions together, robust recalibration is likely to improve accuracy in most wrong-sided decision problems. Robust recalibration is strictly preferable in particular for questions where a binary truth (conditional on the state) exists and extremizing adjustments cannot overshoot the true probability. In problems where the average forecast is not wrong-sided, relative performance depends on the size of the extremizing adjustment, which is determined by how the prior prediction compares to 0.5. We may expect similar performance to simple extremization when estimated priors are in the vicinity of the uninformative prior.

Before continuing to the empirical section of the paper, it is useful to discuss how we have set the tuning parameter γ in our empirical analysis. Recall that γ controls the intensity of extremization away from the estimated prior. As shown in Figure 1, the expected prediction in states $\{\omega_B, \omega_G\}$ satisfies $b < E[P|\omega_B] < P(s_\emptyset) < E[P|\omega_G] < g$. Perfect calibration is achieved when extremization away from $P(s_\emptyset)$ is such that the transformed probability is b

in state ω_B and g in state ω_G. The optimal value of γ depends on the level of conservatism in the average prediction and informativeness of the prior prediction. To illustrate, suppose the actual state is ω_G. Given $P(s_\emptyset) < E[P|\omega_G] < g$, optimal γ is lower if $P(s_\emptyset)$ is closer to g. In contrast, optimal γ would be higher if the prior is biased towards b.

Robust recalibration does not know the optimal value of γ as b and g are unknown, and additional information (such as past data) that may allow estimation of γ is assumed to be unavailable within a single-question aggregation problem. In what follows, we present a wide range of values of γ to investigate how sensitive our approach is to the tuning parameter. Further, when making performance comparisons to other single-question algorithms, we have restricted attention to the tuning parameter range suggested in Baron et al. (2014) and show that our algorithm outperforms the others for both the largest and smallest parameter in this range.

Section 5 tests the robust recalibration method $t_{RR}(\bar{P})$ 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_{RR} where all decision problems are considered unbiased. Thus, we will consider $t_{RR}(\bar{P})$ 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 then compare t_{RR} with various single-question 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.[7]

5.1 Data Sets

We investigate the empirical performance of robust recalibration using four distinct types of experimental tasks taken from Wilkening et al. (2022) and Howe et al. (2024). Appendix C provides example questions from each data set.

The first set of data consists of simple true/false scientific statements. For each statement, participants report a probabilistic prediction on the statement being true as well as a meta-prediction on the average of other participants' predictions. Wilkening et al. (2022) collected data from 500 such statements while Howe et al. (2024) replicated the experiment using a subset of these statements. Each implementation recruited a new sample of participants. 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 participants in a task varied between 79 and 98.

The second data set, referred to as "States" data, was also collected by Wilkening et al. (2022). Each task presented a statement on the largest city of a U.S. state being the capital city of the corresponding state. As seen in Prelec et al. (2017), many people erroneously predict that the largest city is highly likely to be the state capital when they do not know the true answer. As such, the dataset is naturally biased towards true. The States data set includes 50 tasks. In each task, a total of 89 subjects reported probabilistic predictions and meta-predictions on the truth of each statement.

Howe et al. (2024) collected predictions and meta-predictions on various other domains and we use their questions related to art and NFL trivia. In the "Artwork" data set, subjects saw a picture of a drawing and were asked to predict how likely it is that the market value

[7] Supplemental material includes the datasets and R scripts to reproduce all results (Neuwirth, 2022; R Core Team, 2023; RStudio Team, 2020; Wickham, 2007; Wickham et al., 2019).

was more than \$10000. Our data includes 40 decision problems that were repeated in two separate experiments to produce 80 total tasks. The sample size for each task varied between 79 and 87 forecasters. The "NFL" domain tasks presented 50 trivia statements about the NFL draft to a US-based subject pool. Similar to the Artwork data, two runs produced 100 tasks in total. The sample size per task was either 98 or 99.

We note that in two tasks of the Science data, the estimated priors used in the robust recalibration algorithm were outside $(0, 1)$. This can be considered as a failure to estimate $P(s_\emptyset)$ accurately. Appendix D provides the estimated meta-prediction functions and reveals that these were questions where almost all forecasters perfectly predicted the correct answer. Thus, it is likely that these are problems where there is very limited amounts of private information regarding the true state and where idiosyncratic noise in meta-predictions played a large role. We exclude these two science tasks from the results in Section 5.3 and discuss the issue as a potential limitation of our approach in Section 6.[8]

Excluding the two science questions, we had a total of 908 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. Thus, a first step in the analysis is to evaluate the extent to which wrong-sidedness is a problem in the data.

As with most practical forecasting problems, we cannot directly observe the correctly calibrated values of g and b in each of our decision problems. Thus, to classify problems as being wrong-sided, we have to make an assumption regarding these values. In this section, we will assume that $b = 0$ and $g = 1$ so that the state corresponds to the true answer. This assumption is based on the fact that the majority of decision problems are questions that

[8] Alternative approaches to dealing with these two observations such as ignoring the bounds on the prior and running the algorithm or using the original prediction do not change the significance of any test in the paper.

have an objectively correct answer that could be known by a very well-informed forecaster. Thus, the true state could potentially be predicted by a forecaster who receives an infinite number of draws from the potential information system. For $b = 0$ and $g = 1$, Propositions 2 and 3 predict that the robust recalibration algorithm achieves higher accuracy than simple extremization in wrong-sided problems, while performance could be comparable in others. Thus, we expect robust recalibration to improve accuracy on average.

Figure 2 shows the number of tasks in each data set where the average prediction is wrong-sided under the above assumption that $b = 0$ and $g = 1$. As seen, the average prediction is wrong-sided in a considerable number of tasks in each of the data sets. Further, wrong-sided averages are more common in false statements in all task types, suggesting that there is a bias towards true in all datasets.

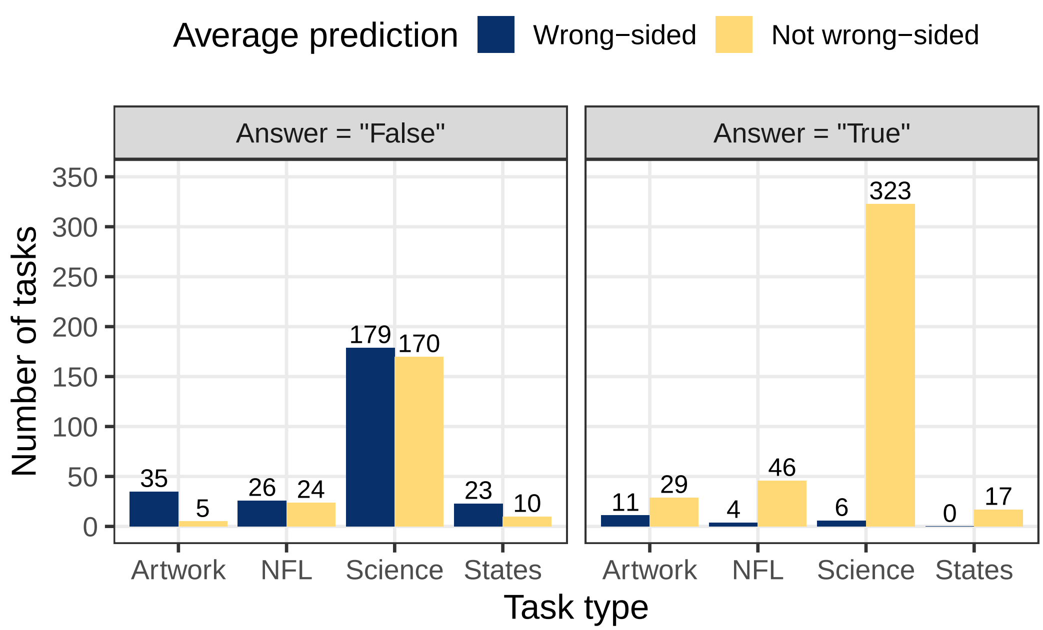


Figure 2: The number of wrong-sided averages in each data set.

Figure 3 estimates the prior using the first stage of our robust recalibration procedure and also supports the conjecture that there is a bias towards true in the data. Estimated priors are typically higher than 0.5. As such, there are likely to be cases where the robust recalibration algorithm transforms an average prediction above 0.5 towards 0 while extremization pushes the same average further towards 1.

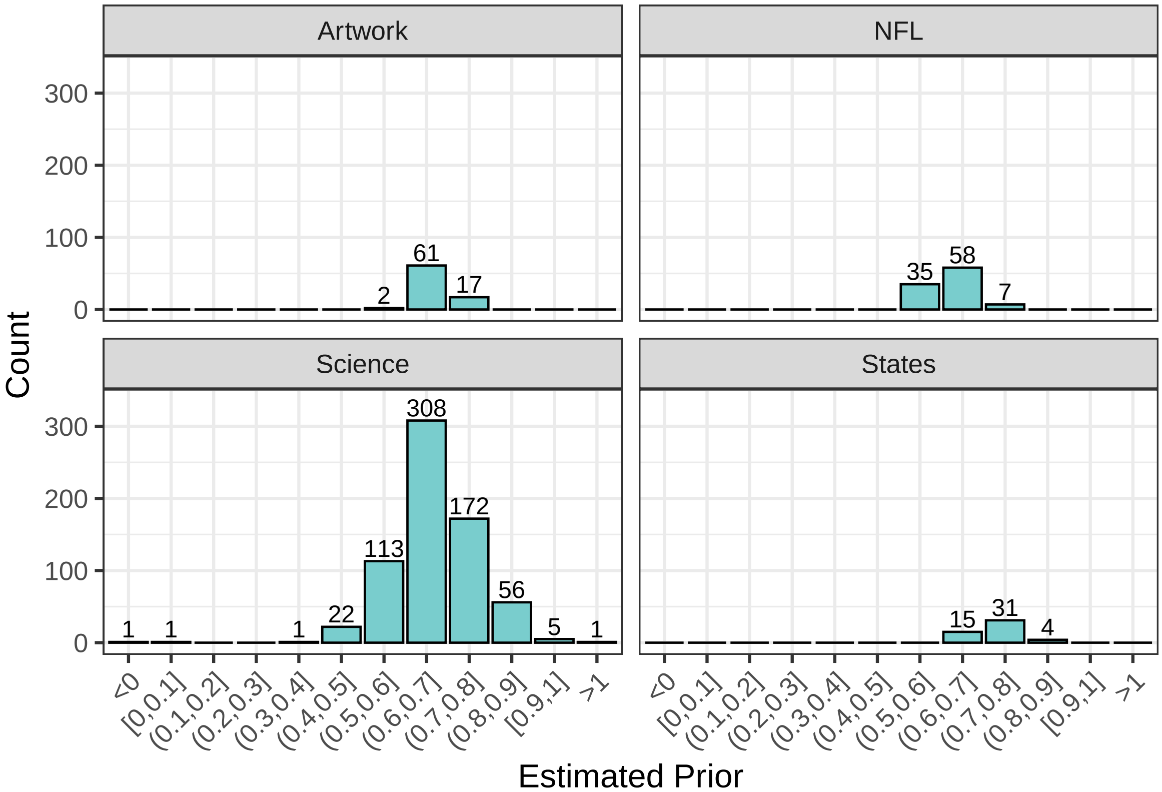


Figure 3: The distribution of estimated priors in each data set.

To understand how the estimated priors influence extremization, we also report the number of decision problems where standard recalibration and robust recalibration procedure recalibrate forecasts towards and away from the true outcome. Tables 1a and 1b show how average predictions compare to 0.5 and the estimated priors respectively. Observations along the diagonal are extremized in the correct direction while observations in the off-diagonal are adjusted in the wrong direction. As can be seen, there are 263 observations in which the average prediction is above 0.5 but the correct answer is false. Of these, the robust recalibration algorithm correctly anti-extremizes 223 observations, while the remaining 40 are still transformed towards 1 as the average prediction is above the estimated prior as well. There are also 415 observations in which the average prediction is above 0.5 and the correct answer is true. Of these, the robust recalibration algorithm incorrectly anti-extremizes 146 observations and the remaining 269 are correctly transformed towards 1. We evaluate how these differences in prediction affect accuracy and calibration in the next section.

(a)

	Correct answer: True	False	Total
$\bar{P} > 0.5$	415	263	678
$\bar{P} < 0.5$	21	209	230
Total	436	472	908

(b)

	Correct answer: True	False	Total
$\bar{P} > \hat{P}(s_\emptyset)$	269	40	309
$\bar{P} < \hat{P}(s_\emptyset)$	167	432	599
Total	436	472	908

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

5.3 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 8 with $\hat{P}(s_\emptyset) = 0.5$, which corresponds to $\delta = 1$. We consider $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ in our implementations of Equation 8 for both extremization and robust recalibration.

Our second analysis compares robust recalibration to various alternative single-question aggregation algorithms that use meta-predictions to improve accuracy. To make comparisons here meaningful, we restrict attention to the range of parameters suggested in Baron et al. (2014) and report results using $\gamma \in \{1.5, 2\}$, which correspond to the suggested lowest and highest values in our reparametrization. We will consider our algorithm as outperforming an alternative if it achieves higher accuracy for both values of γ considered.

The main text reports the analysis when all 908 tasks are used as the basis of the analysis. We provide summary statistic tables for the figures provided in the main text in Appendix E. We also provide an alternative analysis where we compare performance for each of the four prediction tasks separately in Appendix F.

5.3.1 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.[9] Lower scores indicate more accurate forecasts. Each row in the 3×6 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 demonstrates 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 for all values of γ in a paired Wilcoxon sign rank test that treats each decision problem as an observation. Table F1 in Appendix F performs pairwise tests separately for each data set and compares standard extremization to simple average of predictions as well. Robust recalibration achieves substantial and significant improvement in the Science and States tasks, while the level of accuracy is similar to standard extremization in the Artwork and NFL trivia tasks.

[9]Summary statistics for this analysis is provided in Appedix E. Additional task-level analysis is available in Appendix F.

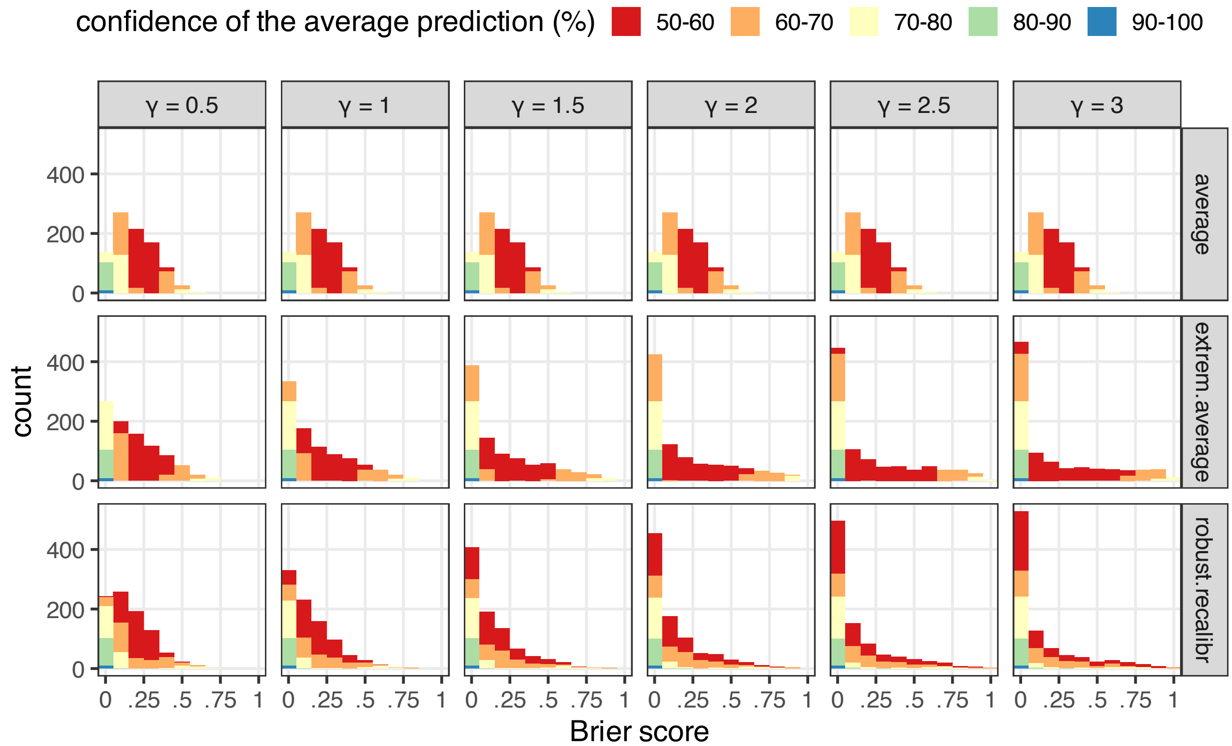


Figure 4: Brier scores of simple average, extremized average and robust-recalibrated probabilities, 908 observations in each panel

γ	Methods		Avg.diff	Med.diff	Test stat.	p-value
0.5	robust.recalibr	extrem.average	-0.0249	-0.0072	V=137,029	<0.0001
1	robust.recalibr	extrem.average	-0.0431	-0.0052	V=143,280	<0.0001
1.5	robust.recalibr	extrem.average	-0.0563	-0.0022	V=148,088	<0.0001
2	robust.recalibr	extrem.average	-0.0658	-0.0008	V=151,761	<0.0001
2.5	robust.recalibr	extrem.average	-0.0728	-0.0003	V=154,699	<0.0001
3	robust.recalibr	extrem.average	-0.0778	-0.0001	V=157,007	<0.0001

Table 2: Two-sided paired Wilcoxon signed rank test of Brier scores, Robust recalibration vs Extremizing away from 0.5. Negative differences indicate higher accuracy for robust recalibration.

Figure 4 also suggests that robust recalibration is particularly effective in transforming low-confidence average predictions. Robust recalibration achieves lower Brier scores when the corresponding average prediction is 50-60% confident, while extremization away from

0.5 leads to higher Brier scores for many such average predictions. Gains in accuracy are especially strong for larger γ. Figure 5 graphs pairwise difference in Brier scores between extremization and robust recalibration. In most tasks where robust recalibration achieves lower Brier scores than simple extremization, the corresponding average prediction is 50-60% confident.

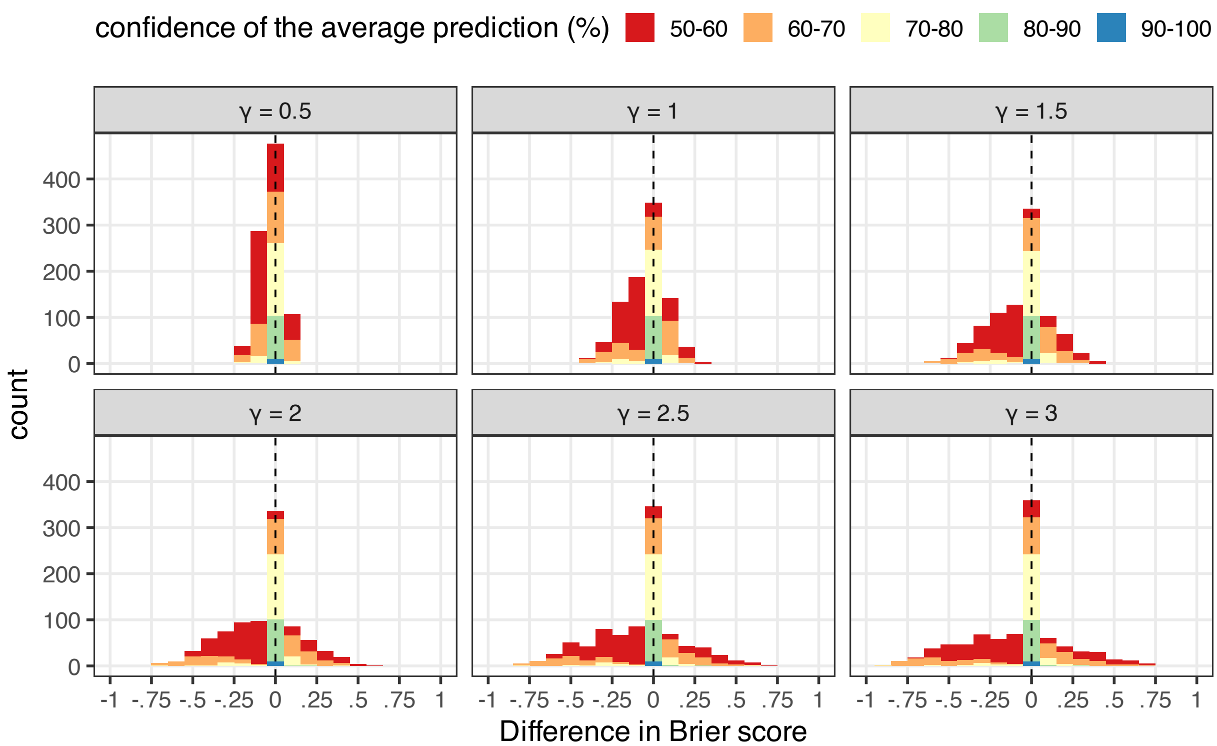


Figure 5: Pairwise differences in Brier score, robust recalibration vs extremized average for $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. Negative differences indicate higher accuracy for robust recalibration.

Why does robust recalibration make the most difference in low-confidence average predictions? Table 3 shows the number of wrong-sided average predictions by confidence across all tasks and reveals that most wrong-sided averages are within the 50-60% confidence category. Recall that wrong-sided averages occur mostly in false statements in our experimental prediction tasks (Table 1) and that estimated priors tend to be above 0.5. As such, simple extremization wrongly transforms these average prediction into high-confidence true predictions. Robust recalibration, by contrast, pushes the average prediction away from the estimated prior instead. This anti-extremization produces better Brier scores on average.

	Confidence of the average prediction (%)					
	50-60	60-70	70-80	80-90	90-100	Total
Wrong-sided	182	85	17	0	0	284
Not wrong-sided	198	160	163	94	9	624
Total	380	245	180	94	9	908

Table 3: Number of wrong-sided average predictions by confidence level.

As we noted in the previous section, robust recalibration also incorrectly anti-extremizes some observations that were true and that had an average prediction above 0.5. Such incorrect recalibrations hurt accuracy relative to the theoretical optimal, but may or may not affect the overall calibration of the algorithm depending on the resulting predicted probabilities. To better understand how well the algorithm calibrates forecasts, we constructed calibration curves 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 6 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 6 suggests that the transformed probabilities from robust recalibration achieve better calibration than standard extremization and the average. In particular for $\gamma \geq 1.5$, 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 6 together imply that the robust recalibration presents a probability transformation that manages to improve both accuracy and calibration.

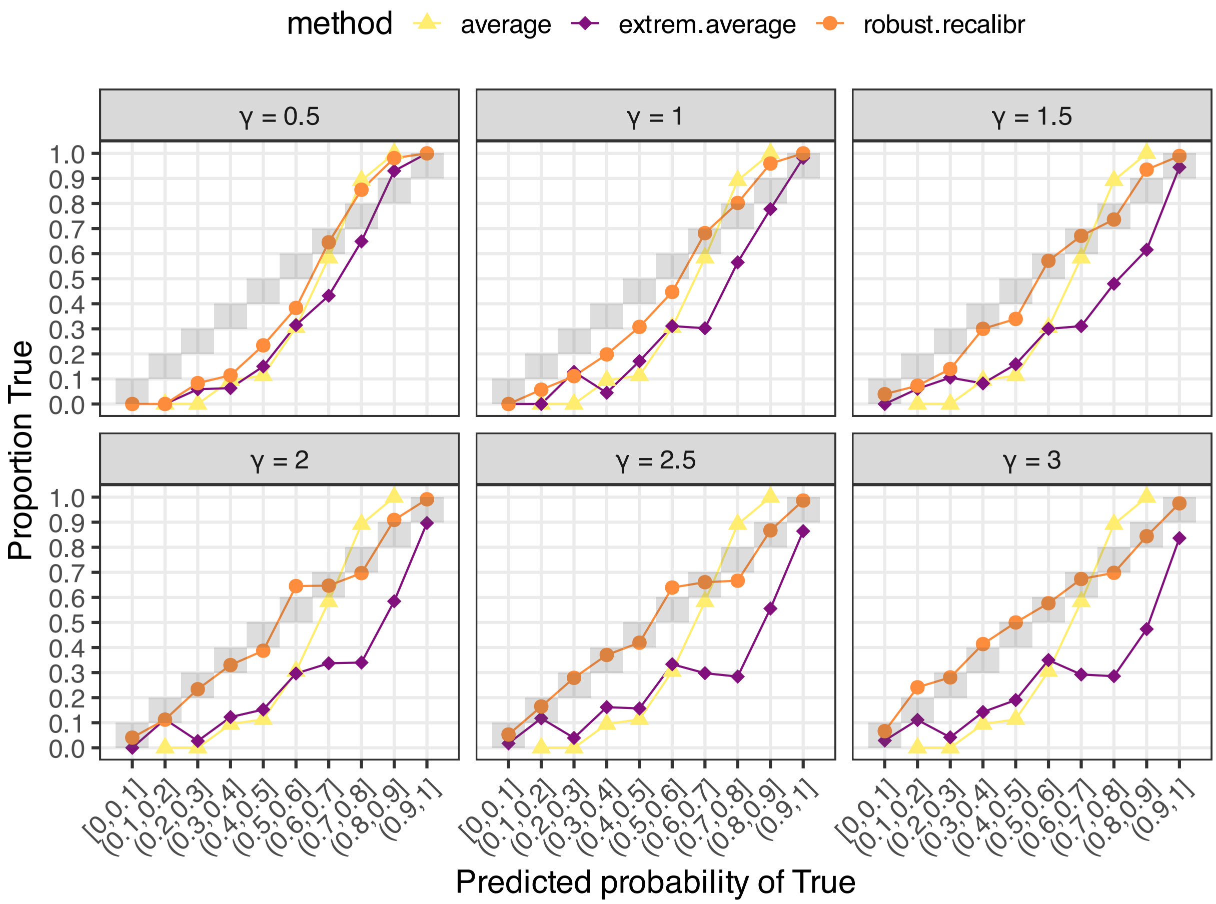


Figure 6: Calibration curves for simple average, extremized average and robust-recalibrated probabilities.

5.3.2 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 estimates the shared and private information in the crowd wisdom, and moves the average away from the shared component. Since shared information contains the prior, correction for the shared-information bias is analogous to an extremization away from the prior and it 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, 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.

As can be seen from the description above, the alternative meta-prediction methods do not have a tuning parameter and thus comparing these algorithms to the robust recalibration method with an extremization parameter that is optimized using a subset of the data is not a fair comparison. To avoid this issue, we instead compare methods using the upper and lower bounds of the parameters that are recommended in the litarature. Baron et al. (2014) estimated that the optimal parameter value in the standard LLO transformation (Equation 2) for the average forecast is between 2.5 and 3, depending on the expertise of forecasters. In our transformation (Equation 7), this would correspond to $\gamma \in [1.5, 2]$, as we define the tuning parameter as $1+\gamma$. When making direct comparisons, we report comparisons using both the lower and upper value in this set and consider the robust recalibration algorithm as an improvement only if it generates an improvement for both of these bounds.[10]

Figure 7 presents the frequency distribution of Brier scores for each of the benchmark algorithms and our robust recalibration method. Panels in the second and third rows show the results for robust recalibration for each $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. 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.

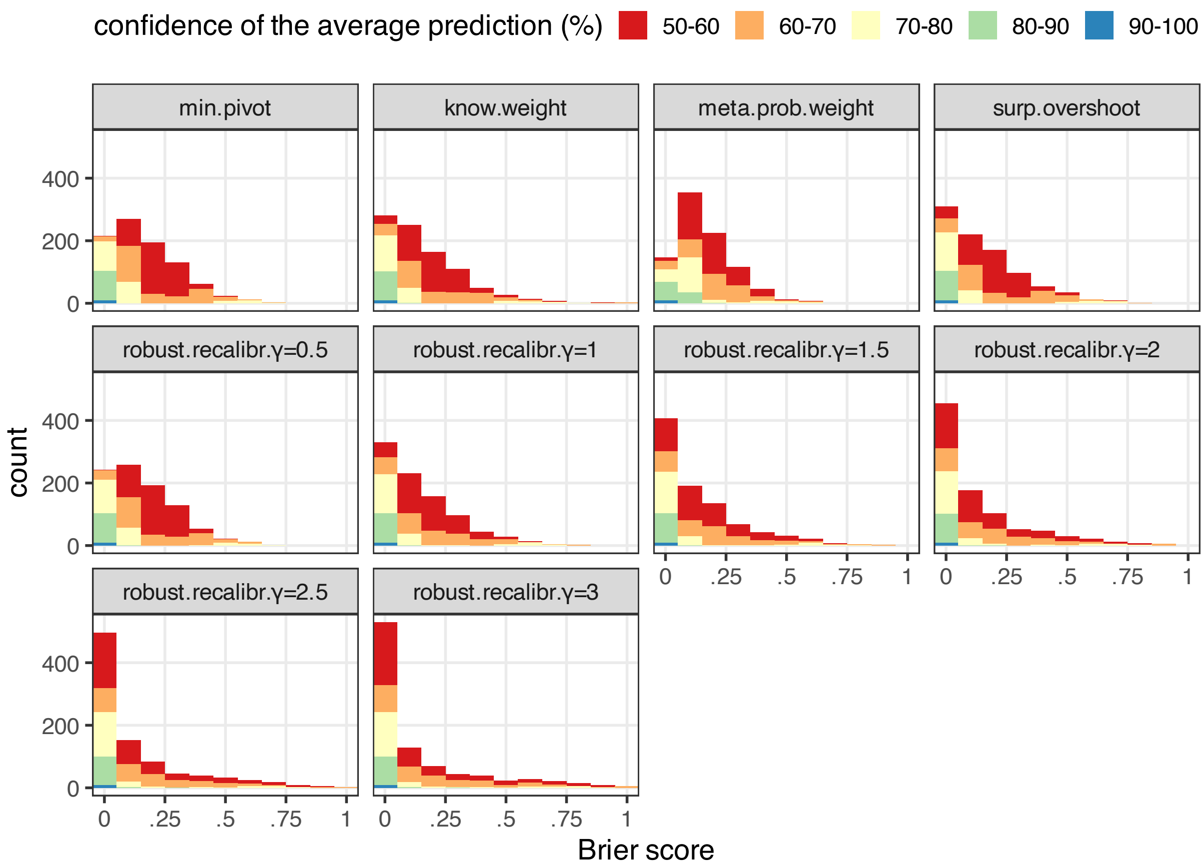


Figure 7: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

Figure 7 demonstrates that robust recalibration achieves very small Brier scores more often than the benchmarks, in particular for $\gamma \geq 1$. The difference between the Brier scores

[10] Table F3 in Appendix F provides comparisons for all $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ for completeness.

of algorithms is significant (ANOVA test, F-value = 5.371, $p < 0.0001$).

We next look at pairwise comparisons of the robust recalibration method with $\gamma \in \{1.5, 2\}$ to the other methods. Table 4 shows that the robust recalibration method achieves higher accuracy against all benchmarks for both values of γ. Table F4 in Appendix F reports the same pairwise tests for each dataset separately. We observe significantly higher accuracy for robust recalibration in the Science and States tasks but find that performance is similar between algorithms in the Arts and NFL trivia tasks. Thus the performance differences between algorithms are likely to relate to characteristics of the underlying data generating process.

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=1.5	know.weight	-0.0230	-0.0150	V=96,184	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	meta.prob.weight	-0.0212	-0.0363	V=103,043	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	min.pivot	-0.0296	-0.0257	V=103,024	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	surp.overshoot	-0.0197	-0.0118	V=123,548	<0.0001	robust.recalibr
robust.recalibr.γ=2	know.weight	-0.0257	-0.0216	V=102,362	<0.0001	robust.recalibr
robust.recalibr.γ=2	meta.prob.weight	-0.0239	-0.0467	V=107,335	<0.0001	robust.recalibr
robust.recalibr.γ=2	min.pivot	-0.0323	-0.0328	V=110,455	<0.0001	robust.recalibr
robust.recalibr.γ=2	surp.overshoot	-0.0224	-0.0188	V=122,617	<0.0001	robust.recalibr

Table 4: Comparison of Brier scores, two-sided paired Wilcoxon signed rank tests, robust recalibration with $\gamma \in \{1.5, 2\}$ vs benchmarks.

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 8 and were constructed using the same methodology as Figure 6. As seen in the diagram, robust recalibration achieves better calibration than the alternatives in most bins for $\gamma \in \{1.5, 2, 2.5, 3\}$. Predicted probabilities of robust-recalibrated aggregates are very close to the actual frequencies. Similar to the results in accuracy above, robust recalibration with sufficiently high γ appears to improve calibration over the alternatives.

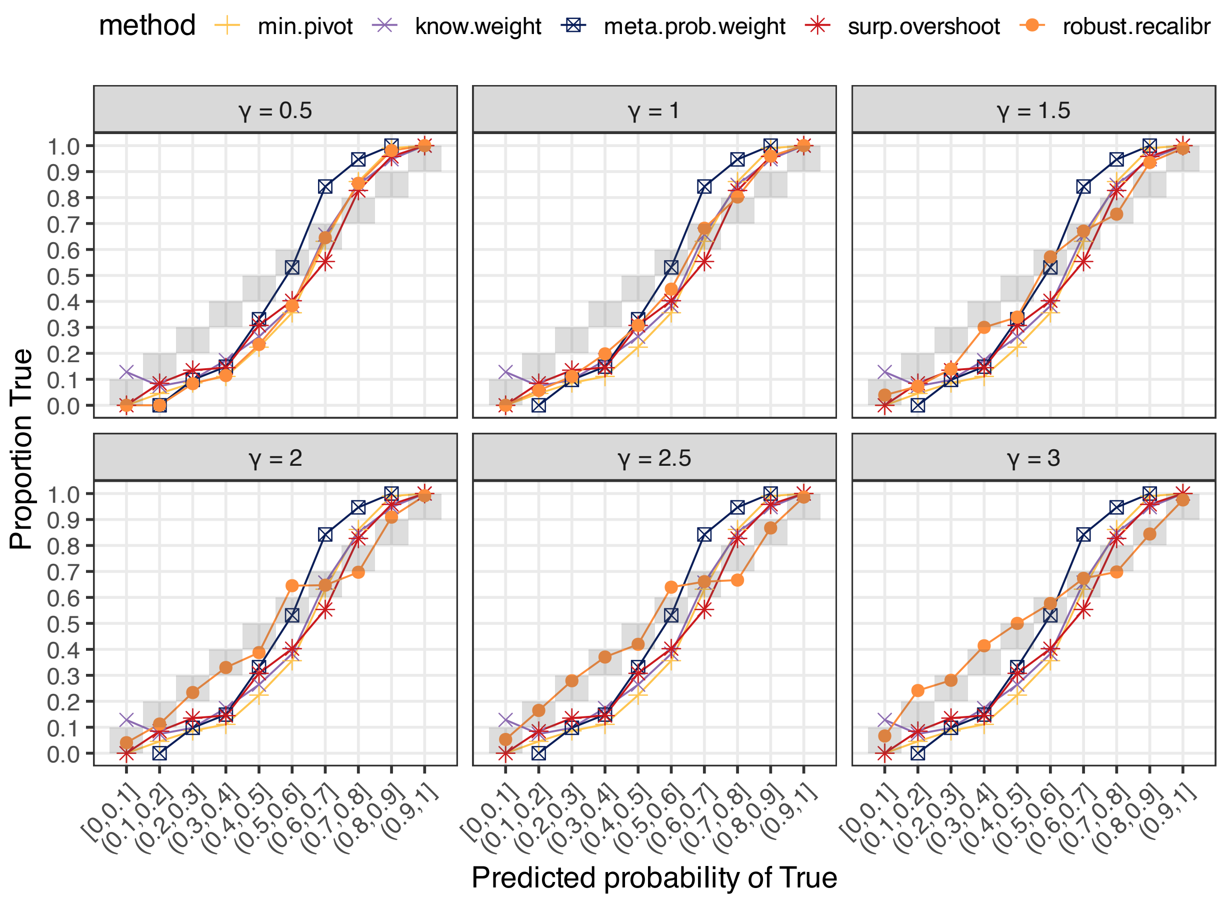


Figure 8: Calibration curves for simple average, extremized average and robust-recalibrated probabilities.

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.

We show that forecasters' meta-beliefs on others' predictions can be used to estimate the prior in single-question forecasting problems. We then 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, which should produce aggregate probability forecasts with better calibration.

We test the performance of robust recalibration using prediction and meta-prediction data from four distinct experimental tasks. We implement robust recalibration with various values of γ, which is a tuning parameter that controls the intensity of extremization away from the estimated prior. Our findings suggest that robust recalibration is effective in improving the accuracy and calibration of probability forecasts. We first demonstrate that robust recalibration outperforms simple extremization away from 0.5 for all values of γ we explored. Robust-recalibrated probabilities achieve lower Brier scores in most tasks and predict the actual frequency of occurrence more accurately than extremized averages. Robust recalibration is particularly effective in transforming wrong-sided averages which are close to 50%, which characterize most wrong-sided averages in our data set. We show that, unlike simple extremization, prior estimation using meta-predictions can detect and transform such wrong-sided averages towards the correct extreme.

We also compared robust recalibration to four single-question aggregation algorithms developed by recent work (Martinie et al., 2020; Palley & Satopää, 2023; Palley & Soll, 2019; Peker, 2023). These algorithms also rely on meta-predictions as well as predictions, but unlike robust recalibration, they do not require a tuning parameter. Thus, they present natural alternatives to our algorithm when meta-prediction data are available. We find that robust recalibration achieves significantly higher accuracy in most tasks when using tuning parameters suggested in the literature. The method also improves calibration provided that γ is sufficiently high. Intuitively, the aggregation algorithms we considered are expected to achieve some improvement in accuracy over simple averaging. Robust recalibration realizes further gains when transformation away from the estimated prior is sufficiently strong, implying that prior estimation is effective in finding the correct direction to transform the average prediction.

Similar to the benchmark algorithms, robust recalibration considers a single forecasting problem where no data other than predictions and meta-predictions are available. Optimal value of γ in a given problem is unknown. Our results suggest that the aggregator may prefer to be aggressive rather than cautious in extremizing away from the estimated prior. Subsequent work may test if this result generalizes to a larger set of forecast aggregation problems. Furthermore, task-level analysis suggests that there is heterogeneity in the relative effectiveness of our algorithm across the tasks studied. Robust recalibration achieved higher accuracy in Science and States tasks, while we see a similar performance to other benchmarks in Artwork and NFL tasks. Future work may investigate if the gains in accuracy differ in various other domains of forecasting as well.

Robust recalibration procedure may have practical limitations due to the prior estimation stage. In two tasks out of 910 in our original data set, the estimated prior probability is not within $(0, 1)$. Appendix D shows that the estimated meta-prediction functions in these two tasks imply meta-predictions outside $(0, 1)$, leading to invalid prior estimates. We observe that in both tasks, predictions are clustered at the correct extreme (0 or 1 depending on the correct answer). In other words, a strong majority of the forecasters were very accurate in their predictions. Robust recalibration uses a linear regression model to estimate the parameters. The actual meta-prediction function may not be estimated accurately when predictions are heavily clustered or the sample of forecasters is small. As discussed in Section 5.2, prior estimation is inaccurate if the estimated meta-prediction function implies meta-predictions outside of the probability scale. Thus, in practical applications, the aggregator can use the information from the estimation procedure to decide on the applicability of robust recalibration.

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147. https://doi.org/10.1037/1076-898X.6.2.130

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

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. https://doi.org/10.1287/deca.2014.0293

Breiman, L. (1996). Stacked regressions. *Machine learning*, *24*, 49–64. https://doi.org/10.1007/BF00117832

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. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3⟨173::AID-BDM261⟩3.0.CO;2-6

Chen, Y.-C., Mueller-Frank, M., & Pai, M. M. (2021). The wisdom of the crowd and higher-order beliefs. https://arxiv.org/abs/2102.02666

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? the surprising informational value of “just asking”. *Judgment and Decision Making*, *14*(2), 135–147. https://doi.org/10.1017/S1930297500003375

Dietrich, F. (2010). Bayesian group belief. *Social choice and welfare*, *35*, 595–626. https://doi.org/10.1007/s00355-010-0453-x

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. https://doi.org/10.1037/0033-295X.101.3.519

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121. https://doi.org/10.1016/j.ijforecast.2012.06.004

Han, Y., & Budescu, D. V. (2022). Recalibrating probabilistic forecasts to improve their accuracy. *Judgment and Decision Making*, *17*(1), 91–123. https://doi.org/10.1017/S1930297500009049

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, *336*(6079), 303–304. https://doi.org/10.1126/science.1221403

Howe, P. D., Martinie, M., & Wilkening, T. (2024). Using cross-domain expertise to aggregate forecasts when within-domain expertise is unknown. *Decision*, *11*(1), 35–59. https://doi.org/10.1037/dec0000212

Jia, Y., Keppo, J., & Satopää, V. (2024). The wisdom of strategically diverse crowds. *Available at SSRN 4855714*. https://ssrn.com/abstract=4855714

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational behavior and human performance*, *21*(1), 61–72. https://doi.org/10.1016/0030-5073(78)90039-9

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362. https://doi.org/10.1126/science.1216549

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

Lee, M. D., & Lee, M. N. (2017). The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, *12*(4), 328–343. https://doi.org/10.1017/S1930297500006227

Libgober, J. (2023). Identifying wisdom (of the crowd): A regression approach. https://arxiv.org/abs/2105.07097

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. https://doi.org/10.1287/opre.2021.2176

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. https://doi.org/10.1371/journal.pone.0232058

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Neuwirth, E. (2022). *Rcolorbrewer: Colorbrewer palettes* [R package version 1.1-3]. https://doi.org/10.32614/CRAN.package.RColorBrewer

Palley, A., & Satopää, V. A. (2023). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Management Science*, *69*(9), 5128–5146. https://doi.org/10.1287/mnsc.2022.4648

Palley, A., & Soll, J. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309. https://doi.org/10.1287/mnsc.2018.3047

Peker, C. (2023). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*, *94*(3), 467–501. https://doi.org/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. https://doi.org/10.1038/nature21054

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(1), 71–91. https://doi.org/10.1111/j.1467-9868.2009.00726.x

Rilling, J. (2024). Neutral pivoting: Strong bias correction for shared information. https://arxiv.org/abs/2404.17737

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Satopää, V. A. (2022). Regularized aggregation of one-off probability predictions. *Operations Research*, *70*(6), 3558–3580. https://doi.org/10.1287/opre.2021.2224

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, *8*(2), 1256–1280. https://doi.org/10.1214/14-AOAS739

Satopää, V. A., Jensen, S. T., Pemantle, R., & Ungar, L. H. (2017). Partial information framework: Model-based aggregation of estimates from diverse information sources.

Electronic Journal of Statistics, *11*(2), 3781–3814. https://doi.org/10.1214/17-EJS1346

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633. https://doi.org/10.1080/01621459.2015.1100621

Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors' probability estimates. *Psychonomic bulletin & review*, *17*(4), 492–498. https://doi.org/10.3758/PBR.17.4.492

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

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. https://doi.org/10.1007/s10994-013-5401-4

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. https://doi.org/10.18637/jss.v021.i12

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/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. https://doi.org/10.1287/mnsc.2020.3919

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl, K. C., & Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, *16*(4), 239–260. https://doi.org/10.1287/deca.2019.0391

Appendices

A Proofs

Lemma 1

This result is due to the fact that the expected posterior prediction generated from an information service is equal to the prediction that would be made at the prior. At the prior:

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

In the main text, we showed that

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

and thus

$$M(s_\emptyset) = q\mathbb{E}[P|\omega_G] + (1-q)\mathbb{E}[P|\omega_B].$$

It follows immediately that $P(s_\emptyset) = M(s_\emptyset)$.

Proposition 1

Consider the case $\omega = \omega_G$. Following the notation in Equation 8, let $t_{RR}(\bar{P})$ denote the robust-recalibrated probability. Also let $t_E(\bar{P})$ be simple-extremized probability ($\delta = 1$ in Equation 8) with the same tuning parameter. Robust recalibration would achieve lower average Brier score if $(t_{RR}(\bar{P}) - g)^2 < (t_E(\bar{P}) - g)^2$, i.e. when the robust-recalibrated probability is more accurate. This expression reduces to $\frac{1}{2}(t_{RR}(\bar{P})+t_E(\bar{P})) < g$, which gives

$$\frac{1}{2}\left(\frac{\delta\bar{P}^{1+\gamma}}{\delta\bar{P}^{1+\gamma}+(1-\bar{P})^{1+\gamma}}+\frac{\bar{P}^{1+\gamma}}{\bar{P}^{1+\gamma}+(1-\bar{P})^{1+\gamma}}\right) < g$$

Note that $\lim_{N\to\infty}\hat{P}(s_\emptyset) = P(s_\emptyset)$, i.e. estimated prior converges to the actual prior prediction at the limit. Thus, $\delta = [(1-P(s_\emptyset))/P(s_\emptyset)]^\gamma$. Also note that $\lim_{N\to\infty}\bar{P} = E[P|\omega_G]$ in state ω_G. Since we consider wrong-sided problems, we have $P(s_\emptyset) < \bar{P} < 0.5$. Then, we have $\delta > 1$ for any γ.

We can define $g'(\delta) = \frac{1}{2}(t_{RR}(\bar{P})+t_E(\bar{P}))$ as the threshold such that, for $g > g'(\delta)$, $t_{RR}(\bar{P})$ is strictly more accurate than $t_E(\bar{P})$ for any $\bar{P}$. Furthermore, $g'(\delta)$ increases as δ increases for all $\delta > 1$. Since δ increases as $P(s_\emptyset)$ decreases, $g'(\delta)$ increases with $|\bar{P}-P(s_\emptyset)|$.

A similar result can be obtained for $\omega = \omega_B$. Robust recalibration is more accurate if $(t_{RR}(\bar{P})-b)^2 < (t_E(\bar{P})-b)^2$ is satisfied, which reduces to $\frac{1}{2}(t_{RR}(\bar{P})+t_E(\bar{P})) > b$. We now have $\lim_{N\to\infty}\bar{P} = E[P|\omega_B]$, $0.5 < \bar{P} < P(s_\emptyset)$, and $\delta < 1$ for any γ. We can define $b'(\delta) = \frac{1}{2}(t_{RR}(\bar{P})+t_E(\bar{P}))$, which decreases as $P(s_\emptyset)$ increases, implying that the threshold $b'(\delta)$ decreases with $|\bar{P}-P(s_\emptyset)|$.

Proposition 2

From the proof of Proposition 1, we know that $\delta > 1$ for $\omega = \omega_G$ and $\delta < 1$ for $\omega = \omega_B$. Then, we simply have $t_E(\bar{P}) < t_{RR}(\bar{P}) < g = 1$ for $\omega = \omega_G$ and $b = 0 < t_{RR}(\bar{P}) < t_E(\bar{P})$. In both states, robust-recalibrated probability is strictly more accurate.

Proposition 3

Consider the case $\omega = \omega_G$. Since average forecast is not wrong sided, we have $0.5 < \bar{P} < 1$. As in the proof of Proposition 1, let $t_{RR}(\bar{P})$ and $t_E(\bar{P})$ denote robust-recalibrated and extremized probabilities. We have $t_E(\bar{P}) < t_{RR}(\bar{P}) < 1 \iff \delta > 1$, which requires $P(s_\emptyset) < 0.5$. Thus, $t_{RR}(\bar{P})$ is strictly more accurate when $|\bar{P}-P(s_\emptyset)| > |\bar{P}-0.5|$. Similarly for $\omega = \omega_B$, we have $0 < t_{RR}(\bar{P}) < t_E(\bar{P}) \iff \delta < 1$, which holds for $P(s_\emptyset) > 0.5$, and $0 < \bar{P} < 0.5$. So, $t_{RR}(\bar{P})$ outperforms $t_E(\bar{P})$ when $|\bar{P} - P(s_\emptyset)| > |\bar{P} - 0.5|$.

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 are exactly two states. This ensured that the algorithm would always identify the correct direction for extremization in large sample. In this section, we use two examples to show that the properties of the algorithm are not guaranteed when there are more than two states. The first example shows that the prediction 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 extremization.

The second example identifies a situation where our algorithm fails to extremize 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.

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 beliefs for each state just as in Section 3.

Example 1: Multiple Cross Points where the estimated posterior is incorrect but the direction of extremization 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.

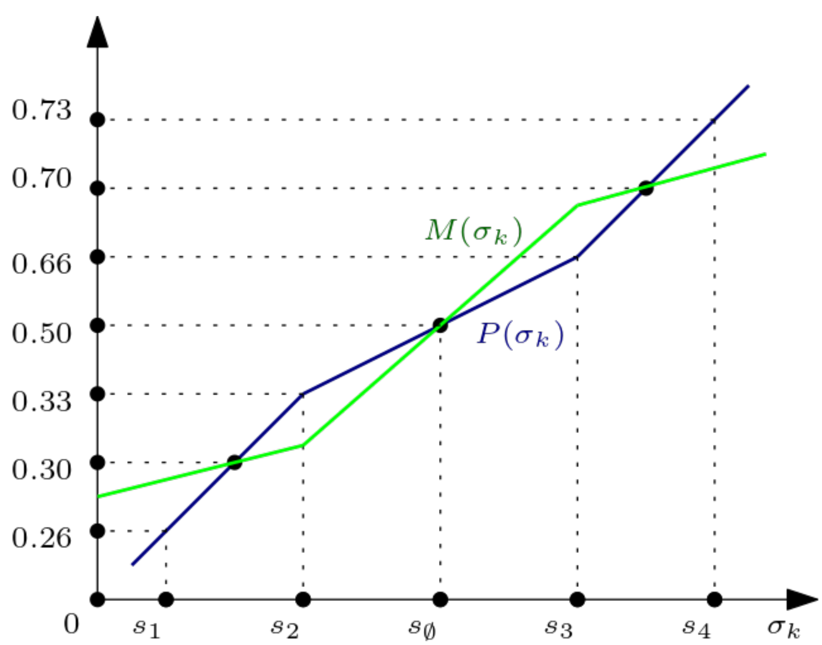


Figure B1: Example 1 prediction and meta-prediction functions (linear extrapolations from the predictions and meta-predictions at $\sigma_k \in \{s_1, s_2, s_\emptyset, s_3, s_4\}$).

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 $P(s_\emptyset)$. 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. Prior prediction is given by $P(s_\emptyset) = \frac{1}{3}0.7 + \frac{1}{3}0.4 + \frac{1}{3}0 = 0.367$. 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 $P(s_1) = 0.21$, $P(s_2) = 0.5$ 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.

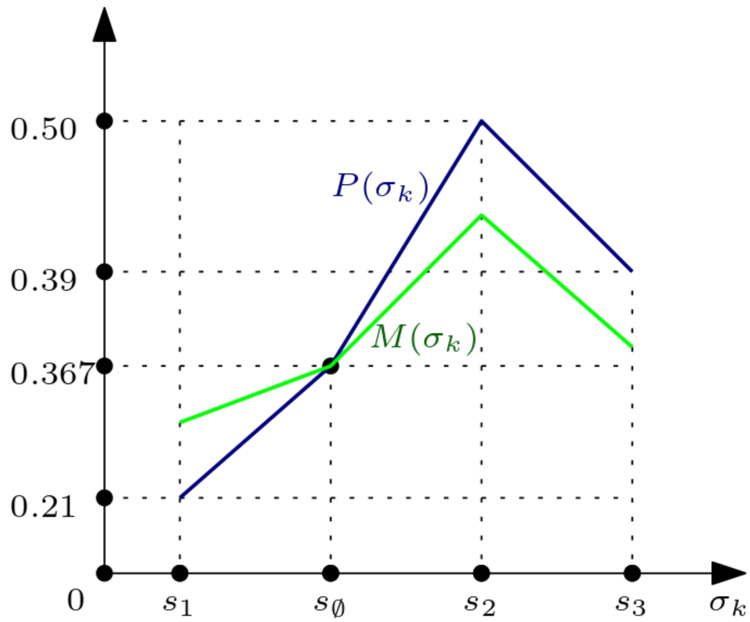


Figure B2: Example 2 prediction and meta-prediction functions

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 4. 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 s_2 being very informative about the predictions of others and the likelihood that the state is not 0. Recall that the posterior beliefs for 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 belief 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.

In the binary framework, signals can be normalized to represent the posterior beliefs on the good state (ω_G). When the true state is ω_G, signals favor a higher probability for the occurrence of E. Therefore, $E[\bar{P}|\omega_G] > E[\bar{P}|\omega_B]$ always holds. 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 one state. Likelihoods in state 0.4 (second row of $\mathbf{Q}$) suggest that all forecasters observe s_2 or s_3, and the corresponding predictions are 0.5 and 0.39. However, in state 0.7 (first row of $\mathbf{Q}$), 30% of forecasters will observe s_1 and predict 0.21. As a result, $E[\bar{P}|state = 0.4] > E[\bar{P}|state = 0.7]$. In other words, the information conveyed by signals 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. Unlike the binary framework, average prediction is higher at a lower state. 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 for the occurrence of the event in a lower state. Thus, we expect robust recalibration to perform well in most applications with more than two states.

C Prediction tasks

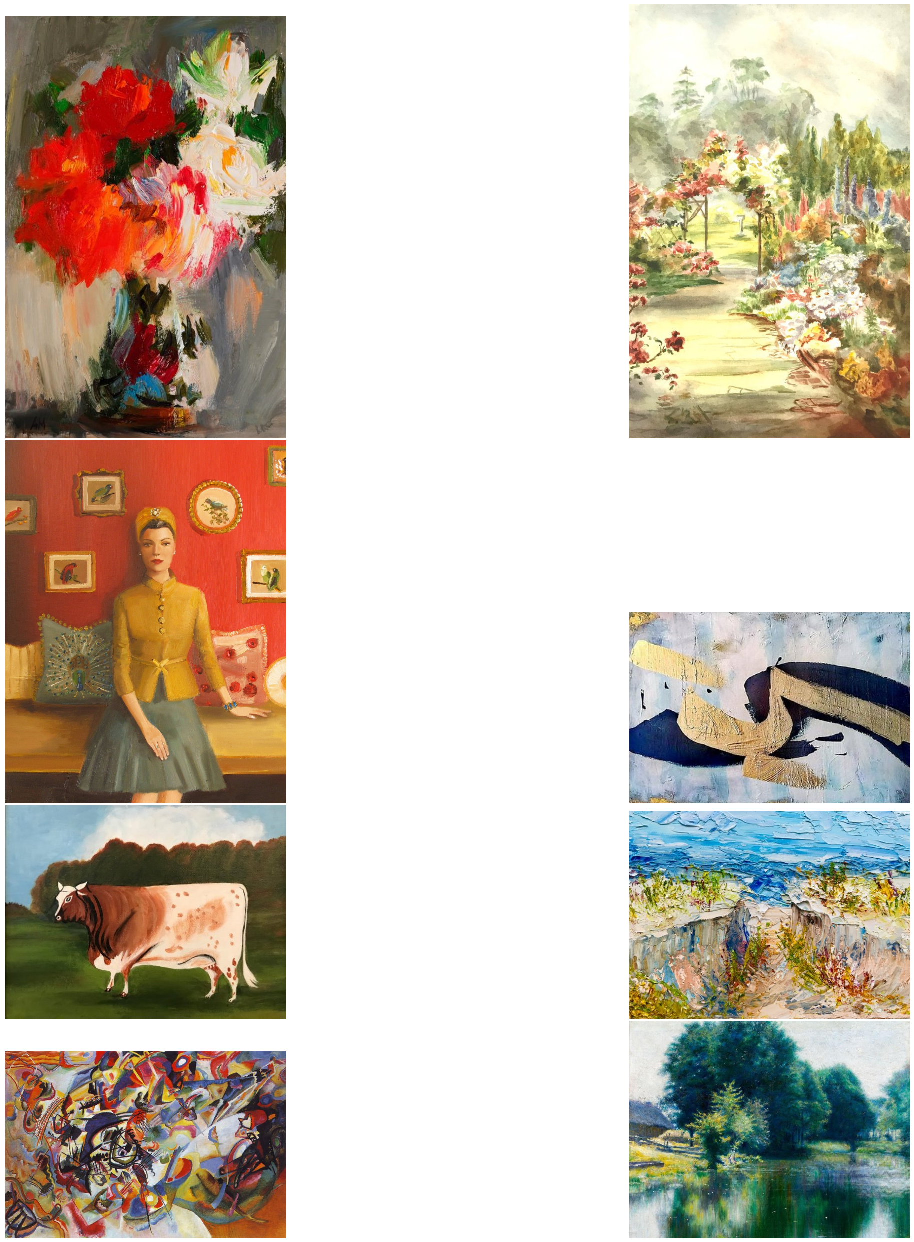
Table C1: Sample statements from Science and States data. See the supplemental material of Wilkening et al. (2022) for full list of statements

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 Two tasks where robust recalibration failed to estimate the prior

Figure D1 shows the estimated meta-prediction function for the two Science tasks where estimated prior lies outside $(0, 1)$. The statements are “Centimetres are a measure of length” and “Fish have fur to keep them warm” with correct answers being true and false respectively.

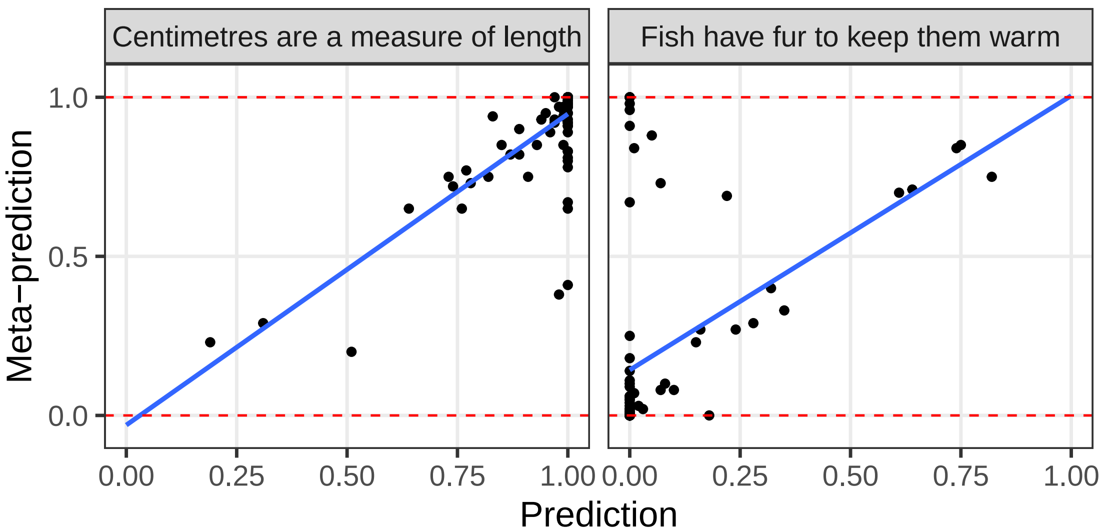


Figure D1: Estimated meta-prediction functions (blue line) in two tasks where estimated prior is not within $(0, 1)$

Estimated meta-prediction functions (as in Equation 6) are $M_k = -0.0302 + 0.9778P_k$ (left panel) and $M_k = 0.1428 + 0.8622P_k$ (right panel). Note that $\hat{\beta}_0 < 0$ for “Centimetres are a measure of length”, which leads to a negative estimated prior of -1.3602 from $\hat{\beta}_0/(1 - \hat{\beta}_1)$. In “Fish have fur to keep them warm”, we have $\hat{\beta}_0 + \hat{\beta}_1 = 1.0049 > 1$, which leads to an estimated prior of 1.0359. Estimated prior probabilities are not within $(0, 1)$.

E Summary statistics and additional figures

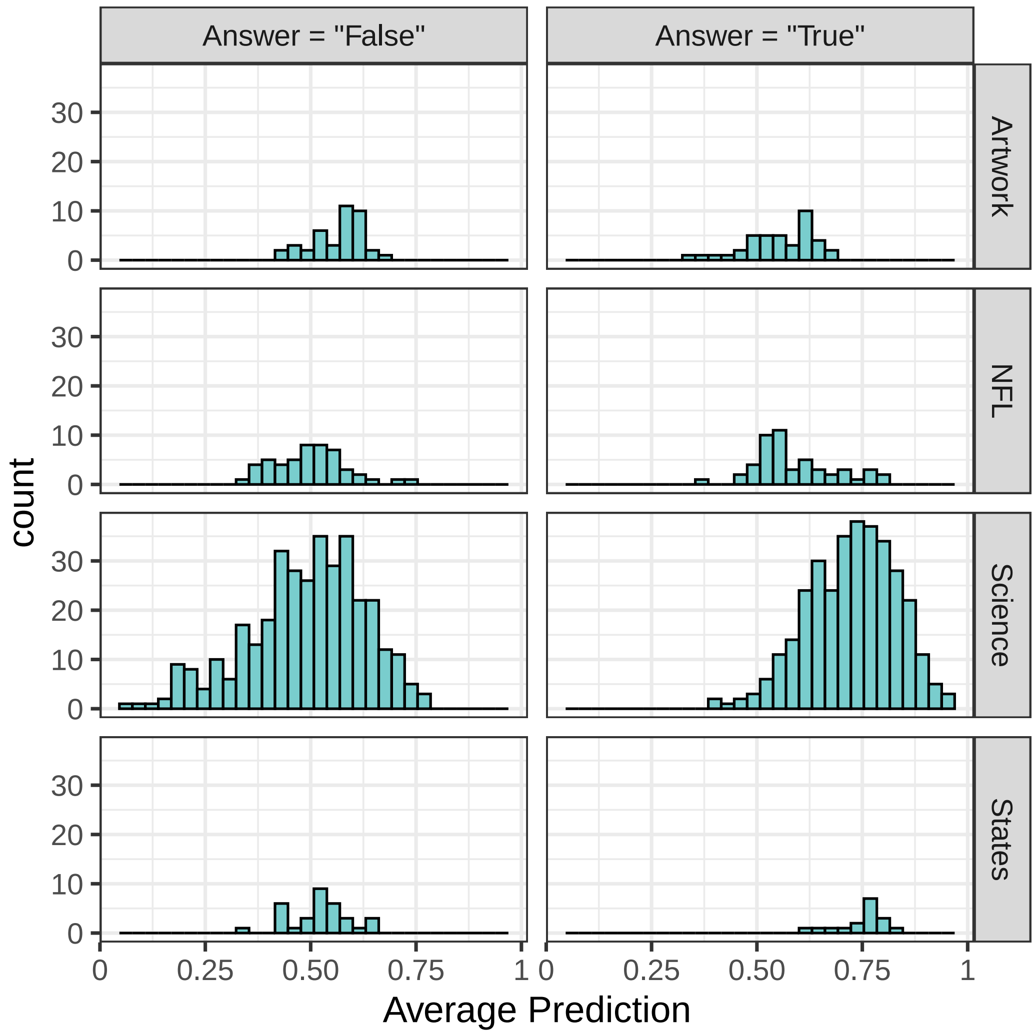


Figure E1: The distribution of average predictions for “True” and “False” statements in each data set.

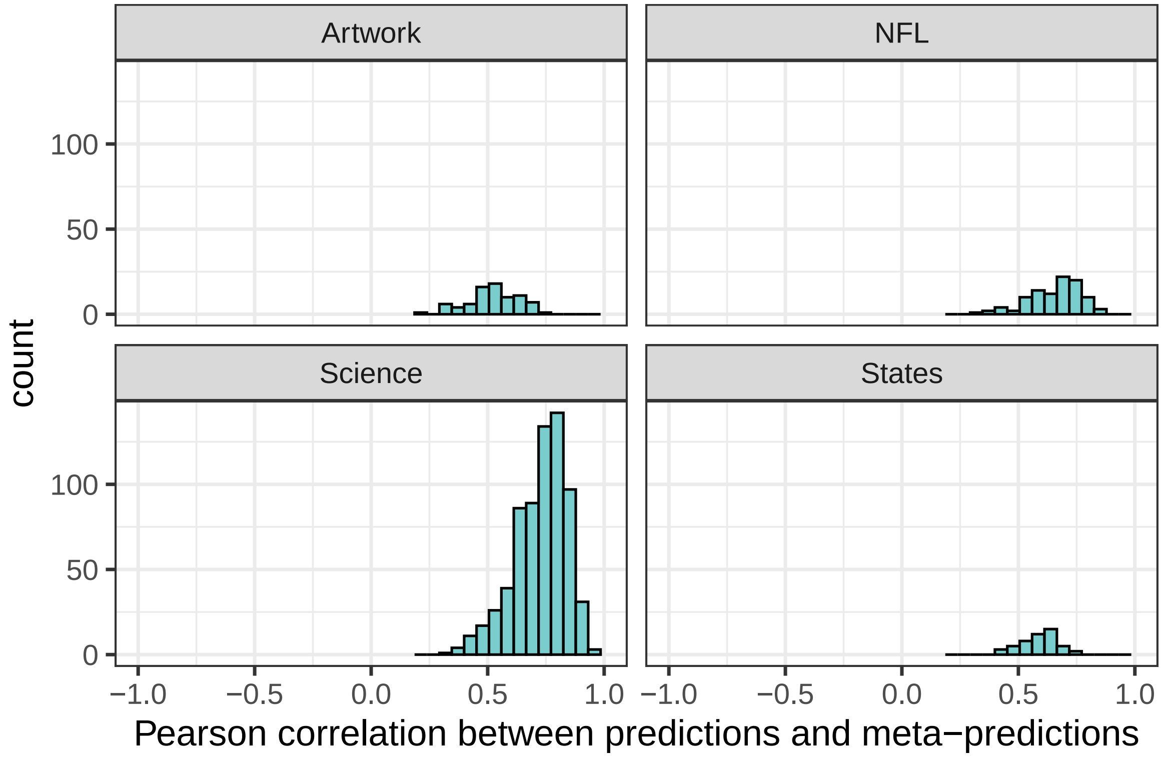


Figure E2: Correlation between predictions and meta-predictions. Each data point represents a task, 910 in total.

method	γ	min	max	mean	lower quartile	median	upper quartile
average		0.0018	0.5878	0.1901	0.0769	0.1737	0.2821
extrem.average	0.5	0.0001	0.7331	0.1859	0.0369	0.1418	0.2987
extrem.average	1	0.0000	0.8376	0.1886	0.0165	0.1143	0.3158
extrem.average	1.5	0.0000	0.9051	0.1944	0.0070	0.0909	0.3332
extrem.average	2	0.0000	0.9459	0.2012	0.0029	0.0715	0.3509
extrem.average	2.5	0.0000	0.9696	0.2083	0.0011	0.0556	0.3688
extrem.average	3	0.0000	0.9831	0.2150	0.0004	0.0428	0.3869
robust.recalibr	0.5	0.0001	0.6529	0.1610	0.0478	0.1314	0.2405
robust.recalibr	1	0.0000	0.7755	0.1455	0.0269	0.0968	0.2224
robust.recalibr	1.5	0.0000	0.8793	0.1381	0.0141	0.0689	0.2037
robust.recalibr	2	0.0000	0.9380	0.1354	0.0068	0.0494	0.1918
robust.recalibr	2.5	0.0000	0.9689	0.1355	0.0031	0.0370	0.1809
robust.recalibr	3	0.0000	0.9846	0.1372	0.0014	0.0259	0.1715

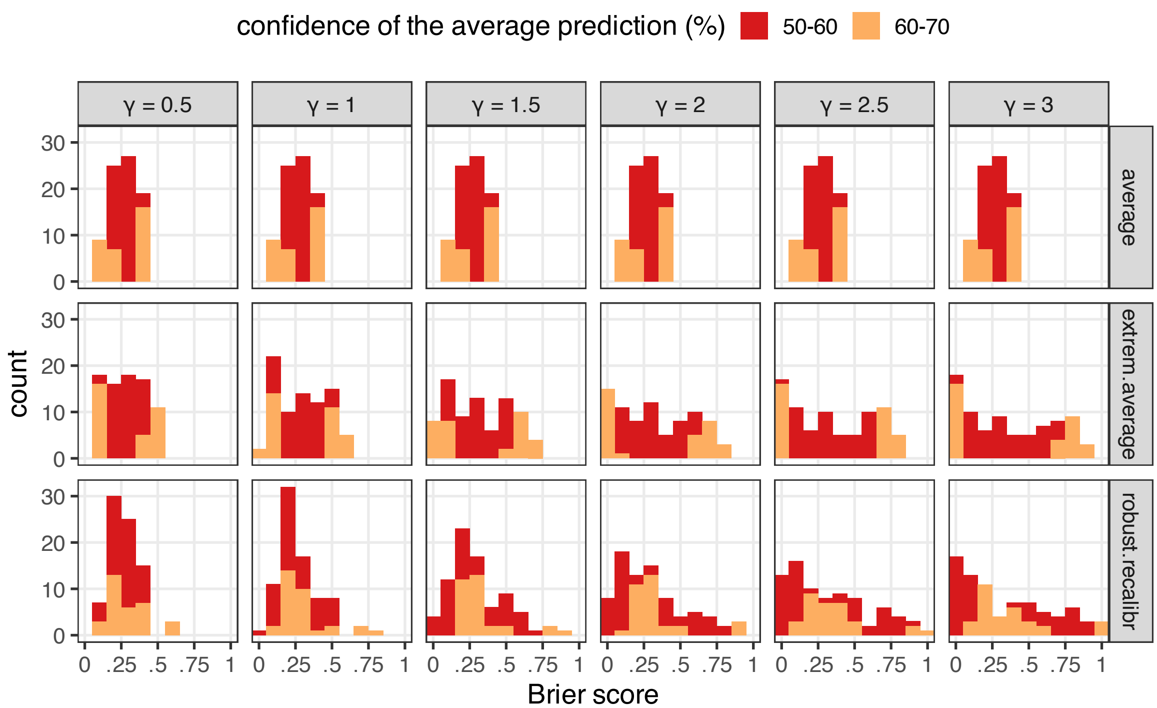
Table E1: Summary statistics, Brier scores in Figure 4.

method	γ	min	max	mean	lower quartile	median	upper quartile
min.pivot		0.0000	0.7031	0.1677	0.0527	0.1399	0.2512
know.weight		0.0000	1.0000	0.1611	0.0366	0.1136	0.2377
meta.prob.weight		0.0014	0.6384	0.1593	0.0723	0.1315	0.2207
surp.overshoot		0.0000	0.7569	0.1578	0.0324	0.1024	0.2500
robust.recalibr	0.5	0.0001	0.6529	0.1610	0.0478	0.1314	0.2405
robust.recalibr	1	0.0000	0.7755	0.1455	0.0269	0.0968	0.2224
robust.recalibr	1.5	0.0000	0.8793	0.1381	0.0141	0.0689	0.2037
robust.recalibr	2	0.0000	0.9380	0.1354	0.0068	0.0494	0.1918
robust.recalibr	2.5	0.0000	0.9689	0.1355	0.0031	0.0370	0.1809
robust.recalibr	3	0.0000	0.9846	0.1372	0.0014	0.0259	0.1715

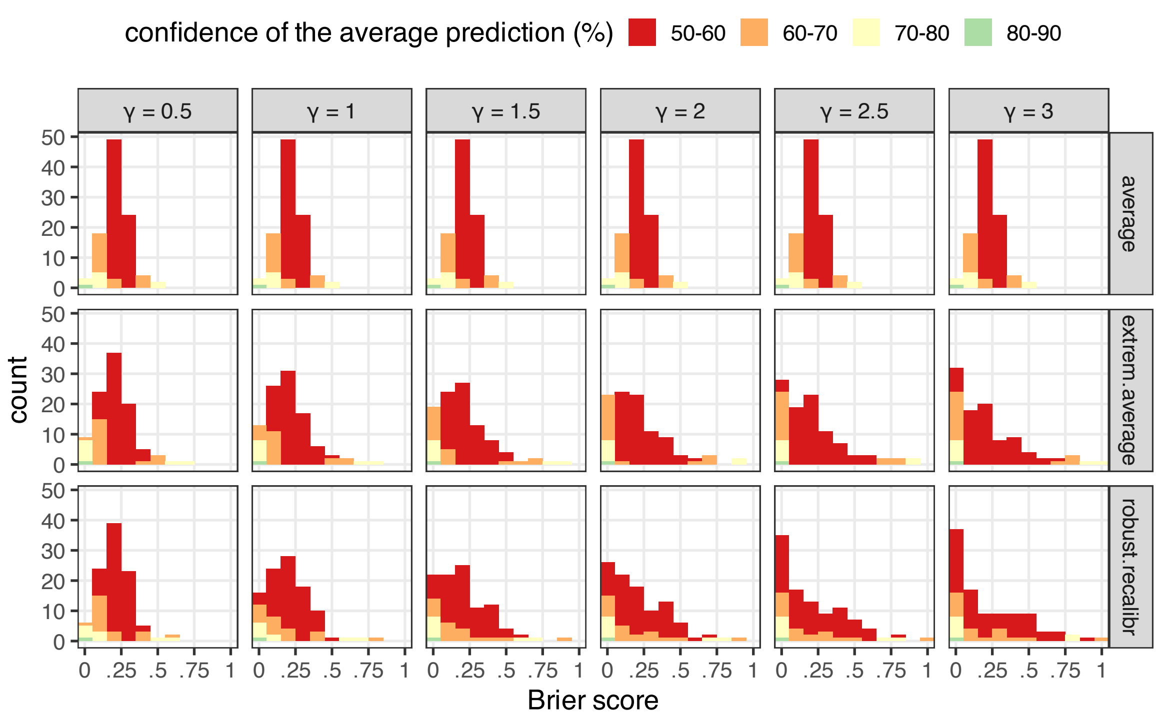
Table E2: Summary statistics, Brier scores in Figure 7.

F Results by data set

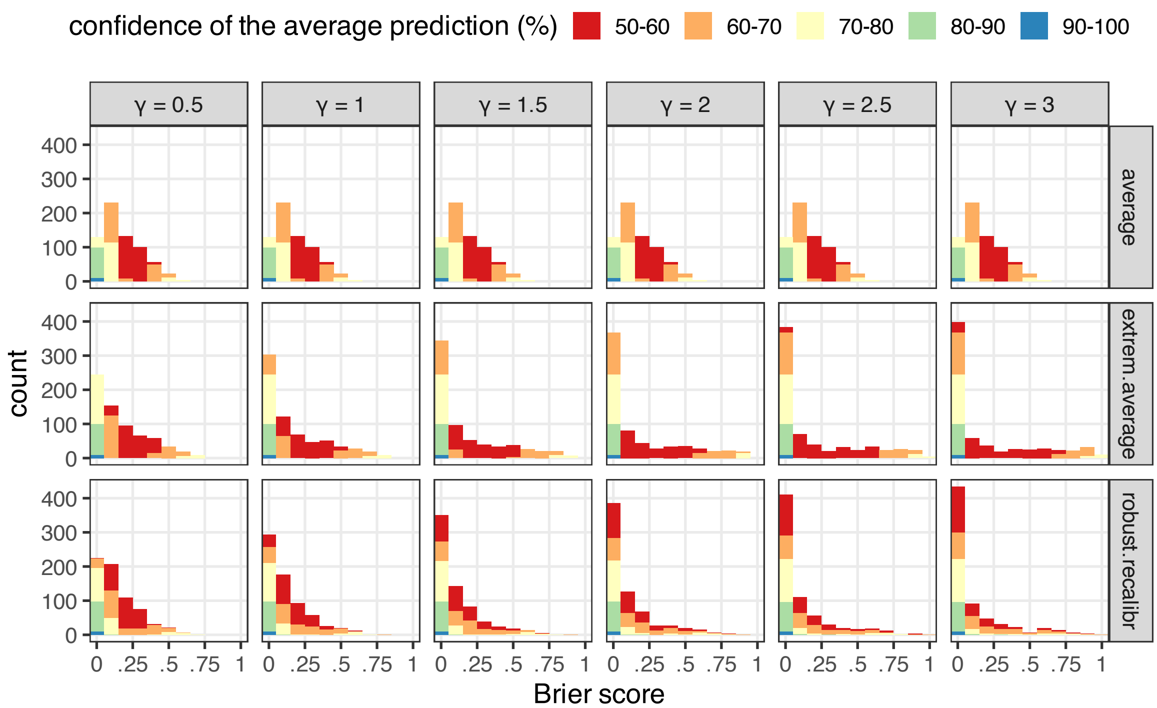
(a) Brier scores, Artwork data only.



(b) Brier scores, NFL data only.



(c) Brier scores, Science data only.



(d) Brier scores, States data only.

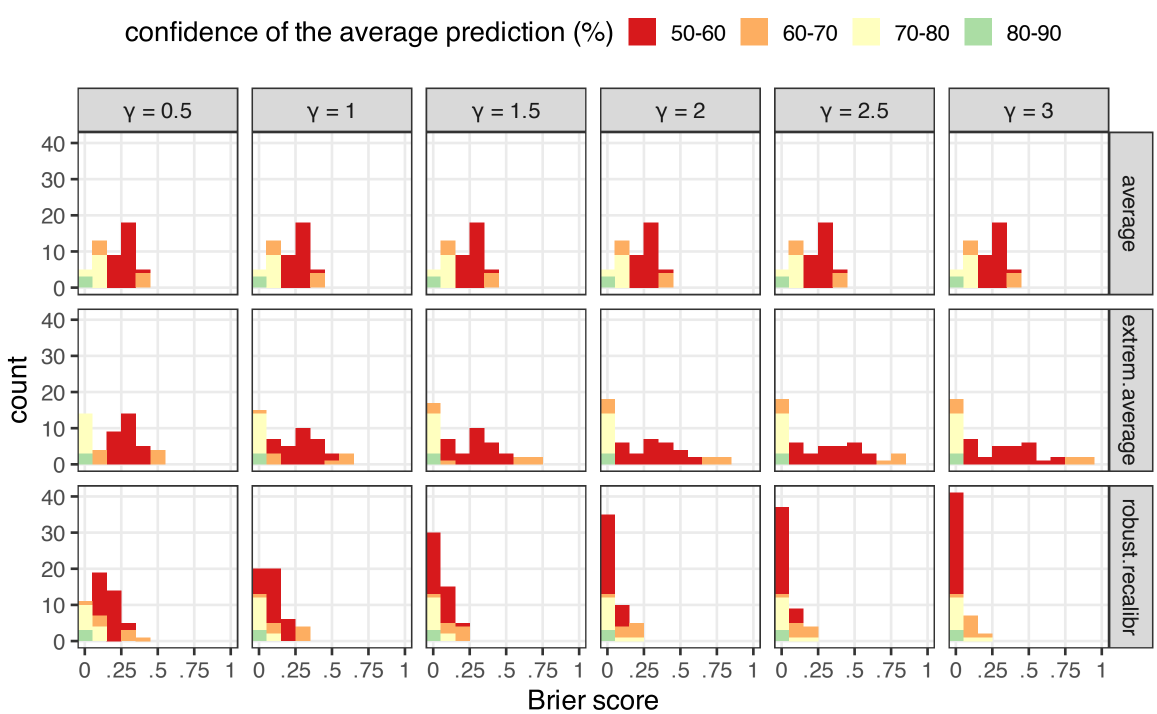
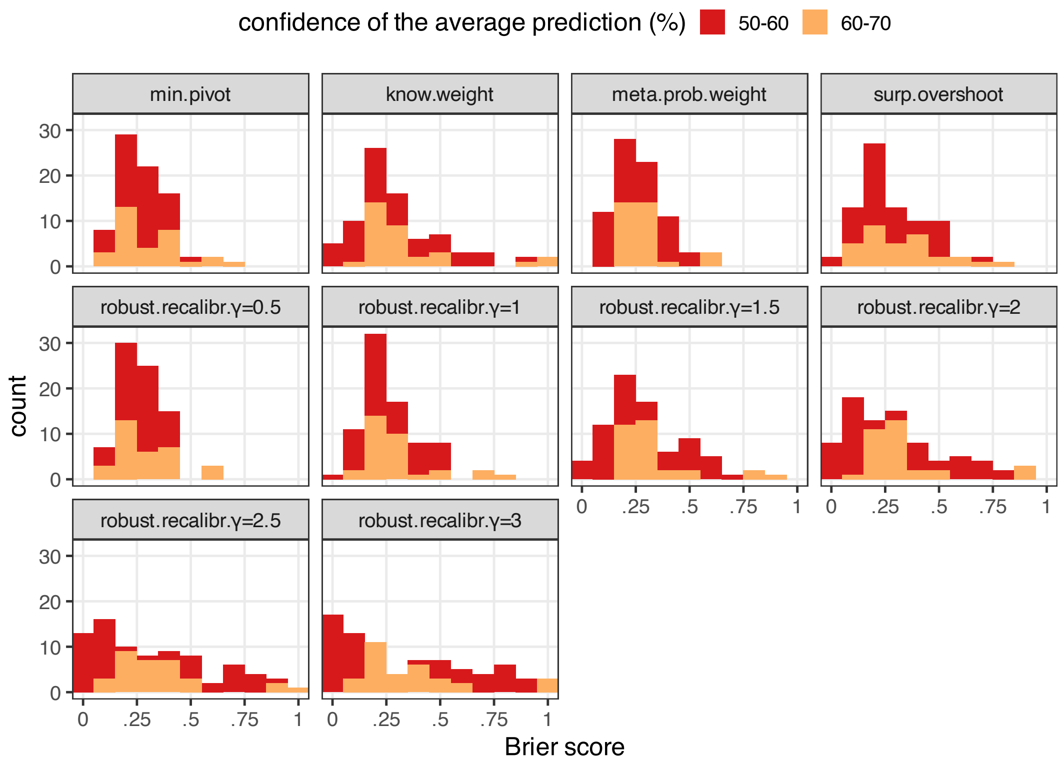
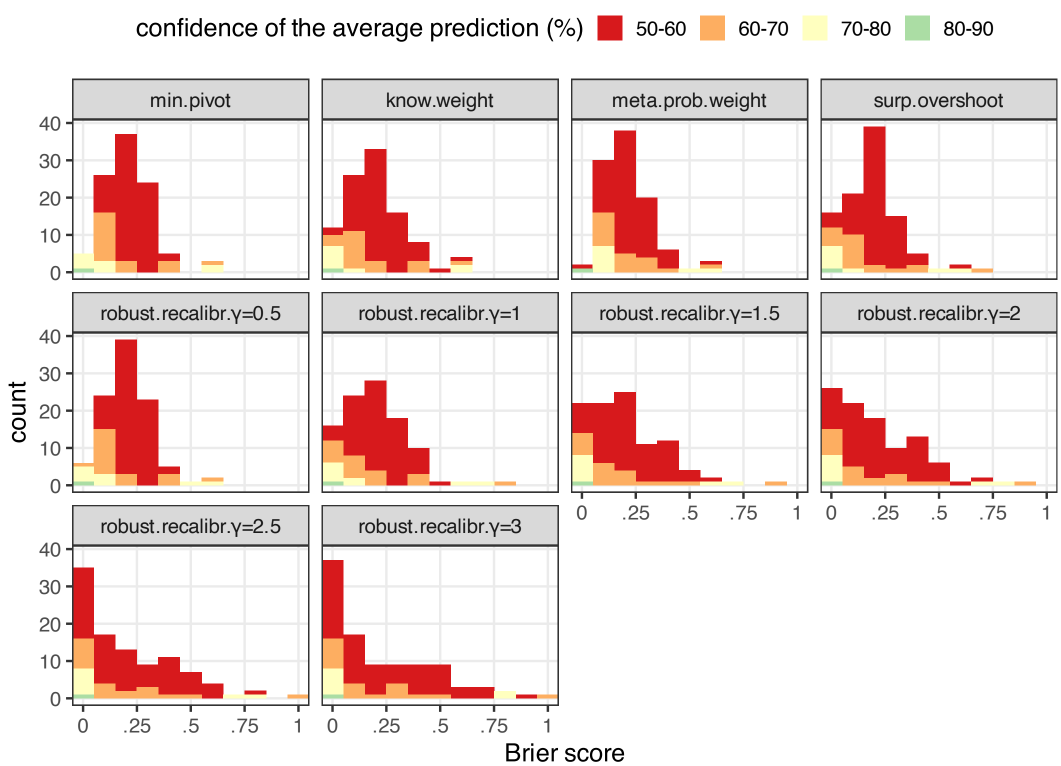


Figure F1: Brier scores of simple average, extremized average and robust-recalibrated probabilities.

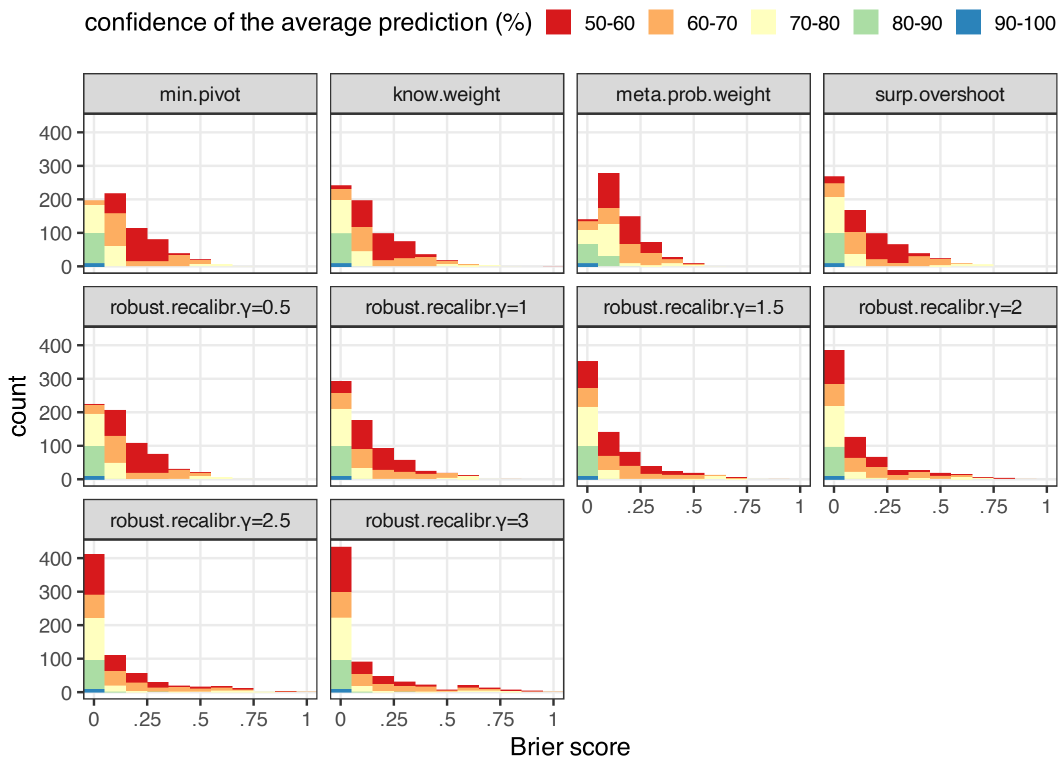
(a) Brier scores, Artwork data only.



(b) Brier scores, NFL data only.



(c) Brier scores, Science data only.



(d) Brier scores, States data only.

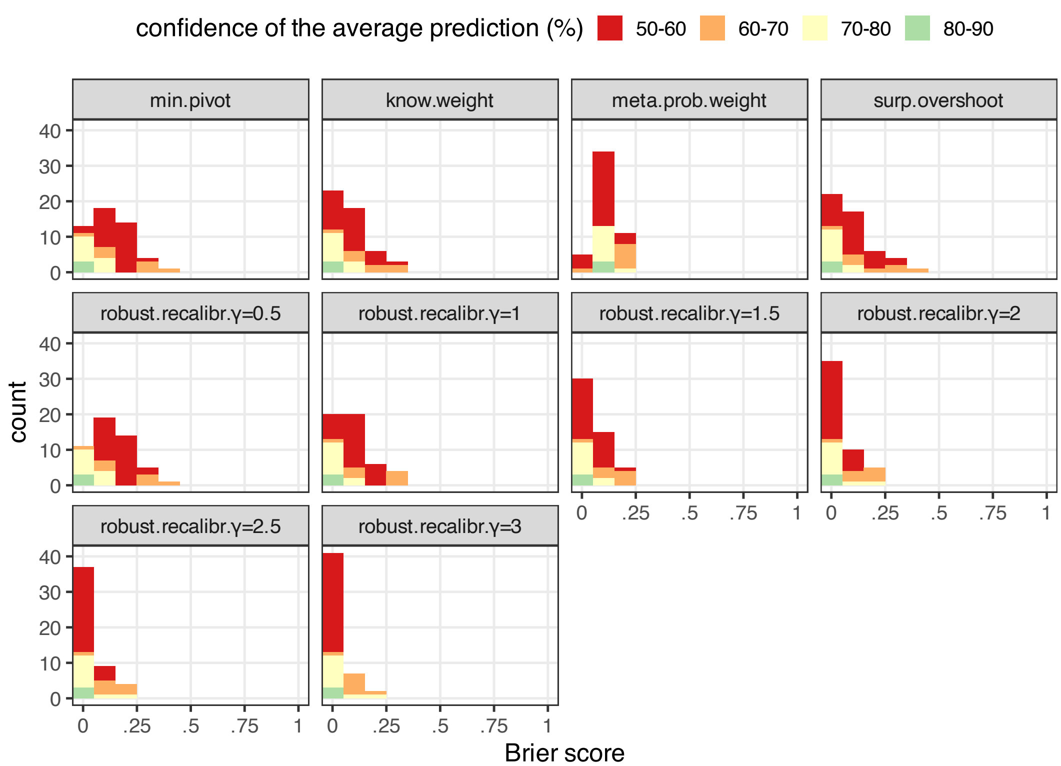


Figure F2: Brier scores of robust recalibration and other benchmarks.

(a) Artwork data only

γ	Method.1	Method.2	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
0.5	extrem.average	average	0.0135	0.0096	V=2,121	0.0164	Method.2
0.5	robust.recalibr	extrem.average	-0.0105	-0.0032	V=1,215	0.0524	No diff.
1	extrem.average	average	0.0292	0.0193	V=2,149	0.0112	Method.2
1	robust.recalibr	extrem.average	-0.0169	0.0021	V=1,261	0.0855	No diff.
1.5	extrem.average	average	0.0460	0.0291	V=2,174	0.0079	Method.2
1.5	robust.recalibr	extrem.average	-0.0206	0.0130	V=1,334	0.1709	No diff.
2	extrem.average	average	0.0630	0.0391	V=2,213	0.0045	Method.2
2	robust.recalibr	extrem.average	-0.0224	0.0265	V=1,379	0.2487	No diff.
2.5	extrem.average	average	0.0795	0.0492	V=2,234	0.0033	Method.2
2.5	robust.recalibr	extrem.average	-0.0232	0.0281	V=1,414	0.3243	No diff.
3	extrem.average	average	0.0951	0.0594	V=2,249	0.0026	Method.2
3	robust.recalibr	extrem.average	-0.0230	0.0212	V=1,446	0.4053	No diff.

(b) NFL data only

γ	Method.1	Method.2	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
0.5	extrem.average	average	-0.0067	-0.0129	V=1,557	0.0009	Method.1
0.5	robust.recalibr	extrem.average	-0.0051	-0.0079	V=2,130	0.1750	No diff.
1	extrem.average	average	-0.0098	-0.0254	V=1,627	0.0020	Method.1
1	robust.recalibr	extrem.average	-0.0062	-0.0097	V=2,303	0.4463	No diff.
1.5	extrem.average	average	-0.0106	-0.0373	V=1,699	0.0045	Method.1
1.5	robust.recalibr	extrem.average	-0.0044	-0.0080	V=2,440	0.7714	No diff.
2	extrem.average	average	-0.0102	-0.0452	V=1,772	0.0097	Method.1
2	robust.recalibr	extrem.average	-0.0007	-0.0055	V=2,508	0.9548	No diff.
2.5	extrem.average	average	-0.0089	-0.0531	V=1,849	0.0202	Method.1
2.5	robust.recalibr	extrem.average	0.0042	-0.0034	V=2,571	0.8757	No diff.
3	extrem.average	average	-0.0072	-0.0622	V=1,900	0.0318	Method.1
3	robust.recalibr	extrem.average	0.0098	-0.0020	V=2,604	0.7872	No diff.

(c) Science data only

γ	Method.1	Method.2	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
0.5	extrem.average	average	-0.0063	-0.0254	V=81,582	<0.0001	Method.1
0.5	robust.recalibr	extrem.average	-0.0264	-0.0050	V=74,929	<0.0001	Method.1
1	extrem.average	average	-0.0045	-0.0377	V=87,242	<0.0001	Method.1
1	robust.recalibr	extrem.average	-0.0461	-0.0024	V=78,104	<0.0001	Method.1
1.5	extrem.average	average	0.0006	-0.0431	V=91,266	<0.0001	Method.1
1.5	robust.recalibr	extrem.average	-0.0608	-0.0007	V=80,416	<0.0001	Method.1
2	extrem.average	average	0.0069	-0.0471	V=94,089	<0.0001	Method.1
2	robust.recalibr	extrem.average	-0.0718	-0.0002	V=82,239	<0.0001	Method.1
2.5	extrem.average	average	0.0134	-0.0489	V=96,155	0.0002	Method.1
2.5	robust.recalibr	extrem.average	-0.0801	-0.0001	V=83,672	<0.0001	Method.1
3	extrem.average	average	0.0195	-0.0510	V=97,698	0.0007	Method.1
3	robust.recalibr	extrem.average	-0.0864	-0.0000	V=84,804	<0.0001	Method.1

(d) States data only

γ	Method.1	Method.2	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
0.5	extrem.average	average	0.0002	-0.0116	V=584	0.6089	No diff.
0.5	robust.recalibr	extrem.average	-0.0667	-0.0808	V=155	<0.0001	Method.1
1	extrem.average	average	0.0071	-0.0224	V=640	0.9846	No diff.
1	robust.recalibr	extrem.average	-0.1183	-0.1256	V=161	<0.0001	Method.1
1.5	extrem.average	average	0.0170	-0.0276	V=688	0.6293	No diff.
1.5	robust.recalibr	extrem.average	-0.1566	-0.1465	V=171	<0.0001	Method.1
2	extrem.average	average	0.0279	-0.0316	V=708	0.4992	No diff.
2	robust.recalibr	extrem.average	-0.1850	-0.1593	V=187	<0.0001	Method.1
2.5	extrem.average	average	0.0388	-0.0350	V=725	0.401	No diff.
2.5	robust.recalibr	extrem.average	-0.2069	-0.1604	V=192	<0.0001	Method.1
3	extrem.average	average	0.0494	-0.0357	V=741	0.3201	No diff.
3	robust.recalibr	extrem.average	-0.2244	-0.1563	V=196	<0.0001	Method.1

Table F1: Two-sided paired Wilcoxon signed rank tests of Brier scores in each data set. Compares robust recalibration, extremizing away from 0.5 and simple average.

Data set	Degrees of Freedom	Mean Sq. Error	F-stat	p-value
Artwork	9	0.0438	1.097	0.362
NFL	9	0.00388	0.142	0.998
Science	9	0.1919	8.125	< 0.0001
States	9	0.07304	13.99	< 0.0001

Table F2: One-way ANOVA test of Brier scores across 10 methods (four benchmark algorithms and robust recalibration with $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$) in each data set. Results suggest significant differences in Science and States data.

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=0.5	know.weight	-0.0001	0.0021	V=247,540	<0.0001	know.weight
robust.recalibr.γ=0.5	meta.prob.weight	0.0017	-0.0075	V=200,532	0.4623	No difference
robust.recalibr.γ=0.5	min.pivot	-0.0067	-0.0017	V=121,239	<0.0001	robust.recalibr
robust.recalibr.γ=0.5	surp.overshoot	0.0032	0.0053	V=246,687	<0.0001	surp.overshoot
robust.recalibr.γ=1	know.weight	-0.0156	-0.0056	V=123,231	<0.0001	robust.recalibr
robust.recalibr.γ=1	meta.prob.weight	-0.0138	-0.0238	V=121,218	<0.0001	robust.recalibr
robust.recalibr.γ=1	min.pivot	-0.0222	-0.0164	V=93,364	<0.0001	robust.recalibr
robust.recalibr.γ=1	surp.overshoot	-0.0123	-0.0047	V=153,070	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	know.weight	-0.0230	-0.0150	V=96,184	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	meta.prob.weight	-0.0212	-0.0363	V=103,043	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	min.pivot	-0.0296	-0.0257	V=103,024	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	surp.overshoot	-0.0197	-0.0118	V=123,548	<0.0001	robust.recalibr
robust.recalibr.γ=2	know.weight	-0.0257	-0.0216	V=102,362	<0.0001	robust.recalibr
robust.recalibr.γ=2	meta.prob.weight	-0.0239	-0.0467	V=107,335	<0.0001	robust.recalibr
robust.recalibr.γ=2	min.pivot	-0.0323	-0.0328	V=110,455	<0.0001	robust.recalibr
robust.recalibr.γ=2	surp.overshoot	-0.0224	-0.0188	V=122,617	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	know.weight	-0.0256	-0.0240	V=110,829	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	meta.prob.weight	-0.0238	-0.0550	V=114,400	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	min.pivot	-0.0322	-0.0383	V=116,401	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	surp.overshoot	-0.0223	-0.0220	V=125,542	<0.0001	robust.recalibr
robust.recalibr.γ=3	know.weight	-0.0239	-0.0274	V=118,513	<0.0001	robust.recalibr
robust.recalibr.γ=3	meta.prob.weight	-0.0221	-0.0588	V=120,723	<0.0001	robust.recalibr
robust.recalibr.γ=3	min.pivot	-0.0305	-0.0421	V=121,302	<0.0001	robust.recalibr
robust.recalibr.γ=3	surp.overshoot	-0.0206	-0.0244	V=129,139	<0.0001	robust.recalibr

Table F3: Comparison of Brier scores, two-sided paired Wilcoxon signed rank tests, robust recalibration with $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ vs benchmarks.

(a) Artwork data only

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=0.5	know.weight	-0.0395	-0.0050	V=1,368	0.2277	No difference
robust.recalibr.γ=0.5	meta.prob.weight	-0.0038	-0.0070	V=1,535	0.6853	No difference
robust.recalibr.γ=0.5	min.pivot	-0.0046	-0.0011	V=1,281	0.1045	No difference
robust.recalibr.γ=0.5	surp.overshoot	-0.0162	-0.0010	V=1,413	0.3220	No difference
robust.recalibr.γ=1	know.weight	-0.0302	-0.0039	V=1,275	0.0985	No difference
robust.recalibr.γ=1	meta.prob.weight	0.0054	-0.0005	V=1,710	0.6677	No difference
robust.recalibr.γ=1	min.pivot	0.0047	0.0070	V=1,645	0.9065	No difference
robust.recalibr.γ=1	surp.overshoot	-0.0069	0.0036	V=1,480	0.5034	No difference
robust.recalibr.γ=1.5	know.weight	-0.0170	-0.0119	V=1,203	0.0458	robust.recalibr
robust.recalibr.γ=1.5	meta.prob.weight	0.0186	-0.0124	V=1,731	0.5961	No difference
robust.recalibr.γ=1.5	min.pivot	0.0178	0.0133	V=1,799	0.3919	No difference
robust.recalibr.γ=1.5	surp.overshoot	0.0062	-0.0010	V=1,718	0.6400	No difference
robust.recalibr.γ=2	know.weight	-0.0019	-0.0289	V=1,387	0.2648	No difference
robust.recalibr.γ=2	meta.prob.weight	0.0337	-0.0051	V=1,845	0.2816	No difference
robust.recalibr.γ=2	min.pivot	0.0329	0.0198	V=1,928	0.1403	No difference
robust.recalibr.γ=2	surp.overshoot	0.0214	-0.0070	V=1,926	0.1428	No difference
robust.recalibr.γ=2.5	know.weight	0.0139	-0.0027	V=1,642	0.9179	No difference
robust.recalibr.γ=2.5	meta.prob.weight	0.0495	-0.0029	V=1,977	0.0873	No difference
robust.recalibr.γ=2.5	min.pivot	0.0487	0.0264	V=2,047	0.0408	min.pivot
robust.recalibr.γ=2.5	surp.overshoot	0.0372	-0.0096	V=2,048	0.0403	robust.recalibr
robust.recalibr.γ=3	know.weight	0.0296	0.0099	V=1,840	0.2924	No difference
robust.recalibr.γ=3	meta.prob.weight	0.0652	-0.0104	V=2,106	0.0199	robust.recalibr
robust.recalibr.γ=3	min.pivot	0.0645	0.0332	V=2,118	0.0170	min.pivot
robust.recalibr.γ=3	surp.overshoot	0.0529	0.0176	V=2,115	0.0177	surp.overshoot

(b) NFL data only

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=0.5	know.weight	-0.0005	0.0030	V=3,060	0.0661	No difference
robust.recalibr.γ=0.5	meta.prob.weight	-0.0014	0.0000	V=2,550	0.9329	No difference
robust.recalibr.γ=0.5	min.pivot	-0.0011	-0.0004	V=2,222	0.2983	No difference
robust.recalibr.γ=0.5	surp.overshoot	0.0083	0.0077	V=3,441	0.0016	No difference
robust.recalibr.γ=1	know.weight	-0.0047	-0.0016	V=2,198	0.2616	No difference
robust.recalibr.γ=1	meta.prob.weight	-0.0056	-0.0132	V=1,933	0.0420	robust.recalibr
robust.recalibr.γ=1	min.pivot	-0.0053	-0.0110	V=1,970	0.0566	No difference
robust.recalibr.γ=1	surp.overshoot	0.0041	0.0003	V=2,673	0.6120	No difference
robust.recalibr.γ=1.5	know.weight	-0.0037	-0.0105	V=1,981	0.0617	No difference
robust.recalibr.γ=1.5	meta.prob.weight	-0.0046	-0.0253	V=2,015	0.0798	No difference
robust.recalibr.γ=1.5	min.pivot	-0.0044	-0.0204	V=2,148	0.1955	No difference
robust.recalibr.γ=1.5	surp.overshoot	0.0050	-0.0062	V=2,445	0.7846	No difference
robust.recalibr.γ=2	know.weight	0.0004	-0.0168	V=2,173	0.2268	No difference
robust.recalibr.γ=2	meta.prob.weight	-0.0004	-0.0402	V=2,210	0.2795	No difference
robust.recalibr.γ=2	min.pivot	-0.0002	-0.0268	V=2,307	0.4546	No difference
robust.recalibr.γ=2	surp.overshoot	0.0092	-0.0119	V=2,472	0.8568	No difference
robust.recalibr.γ=2.5	know.weight	0.0066	-0.0218	V=2,319	0.4798	No difference
robust.recalibr.γ=2.5	meta.prob.weight	0.0057	-0.0511	V=2,332	0.5080	No difference
robust.recalibr.γ=2.5	min.pivot	0.0060	-0.0291	V=2,415	0.7065	No difference
robust.recalibr.γ=2.5	surp.overshoot	0.0153	-0.0158	V=2,518	0.9822	No difference
robust.recalibr.γ=3	know.weight	0.0139	-0.0250	V=2,454	0.8085	No difference
robust.recalibr.γ=3	meta.prob.weight	0.0130	-0.0558	V=2,454	0.8085	No difference
robust.recalibr.γ=3	min.pivot	0.0133	-0.0313	V=2,517	0.9794	No difference
robust.recalibr.γ=3	surp.overshoot	0.0227	-0.0191	V=2,586	0.8352	No difference

(c) Science data only

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=0.5	know.weight	0.0005	0.0014	V=135,238	<0.0001	know.weight
robust.recalibr.γ=0.5	meta.prob.weight	0.0005	-0.0087	V=105,406	0.0577	No difference
robust.recalibr.γ=0.5	min.pivot	-0.0084	-0.0024	V=55,092	<0.0001	robust.recalibr
robust.recalibr.γ=0.5	surp.overshoot	0.0017	0.0045	V=133,503	0.0003	surp.overshoot
robust.recalibr.γ=1	know.weight	-0.0174	-0.0068	V=53,859	<0.0001	robust.recalibr
robust.recalibr.γ=1	meta.prob.weight	-0.0175	-0.0272	V=57,205	<0.0001	robust.recalibr
robust.recalibr.γ=1	min.pivot	-0.0264	-0.0166	V=39,850	<0.0001	robust.recalibr
robust.recalibr.γ=1	surp.overshoot	-0.0163	-0.0058	V=73,182	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	know.weight	-0.0269	-0.0162	V=43,809	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	meta.prob.weight	-0.0270	-0.0389	V=47,981	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	min.pivot	-0.0359	-0.0253	V=43,628	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	surp.overshoot	-0.0258	-0.0123	V=55,148	<0.0001	robust.recalibr
robust.recalibr.γ=2	know.weight	-0.0316	-0.0216	V=46,463	<0.0001	robust.recalibr
robust.recalibr.γ=2	meta.prob.weight	-0.0317	-0.0481	V=48,503	<0.0001	robust.recalibr
robust.recalibr.γ=2	min.pivot	-0.0406	-0.0327	V=46,822	<0.0001	robust.recalibr
robust.recalibr.γ=2	surp.overshoot	-0.0305	-0.0192	V=54,264	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	know.weight	-0.0334	-0.0244	V=49,251	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	meta.prob.weight	-0.0335	-0.0557	V=50472	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	min.pivot	-0.0424	-0.0378	V=49,365	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	surp.overshoot	-0.0323	-0.0225	V=55,183	<0.0001	robust.recalibr
robust.recalibr.γ=3	know.weight	-0.0336	-0.0278	V=51,837	<0.0001	robust.recalibr
robust.recalibr.γ=3	meta.prob.weight	-0.0337	-0.0576	V=52,322	<0.0001	robust.recalibr
robust.recalibr.γ=3	min.pivot	-0.0426	-0.0416	V=51,598	<0.0001	robust.recalibr
robust.recalibr.γ=3	surp.overshoot	-0.0325	-0.0254	V=56,356	<0.0001	robust.recalibr

(d) States data only

Method	Benchmark	Avg.diff	Med.diff	Test stat.	p-value	Signif. better?
robust.recalibr.γ=0.5	know.weight	0.0551	0.0463	V=1,246	<0.0001	know.weight
robust.recalibr.γ=0.5	meta.prob.weight	0.0337	0.0322	V=932	0.0045	meta.prob.weight
robust.recalibr.γ=0.5	min.pivot	0.0019	0.0008	V=798	0.1225	No difference
robust.recalibr.γ=0.5	surp.overshoot	0.0448	0.0210	V=1,167	<0.0001	surp.overshoot
robust.recalibr.γ=1	know.weight	0.0104	0.0039	V=911	0.0084	know.weight
robust.recalibr.γ=1	meta.prob.weight	-0.0110	-0.0182	V=417	0.0337	robust.recalibr
robust.recalibr.γ=1	min.pivot	-0.0429	-0.0537	V=44	<0.0001	robust.recalibr
robust.recalibr.γ=1	surp.overshoot	0.0001	0.0071	V=696	0.5756	No difference
robust.recalibr.γ=1.5	know.weight	-0.0180	-0.0124	V=273	0.0004	robust.recalibr
robust.recalibr.γ=1.5	meta.prob.weight	-0.0394	-0.0419	V=84	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	min.pivot	-0.0712	-0.0868	V=46	<0.0001	robust.recalibr
robust.recalibr.γ=1.5	surp.overshoot	-0.0283	-0.0132	V=318	0.0021	robust.recalibr
robust.recalibr.γ=2	know.weight	-0.0356	-0.0272	V=138	<0.0001	robust.recalibr
robust.recalibr.γ=2	meta.prob.weight	-0.0570	-0.0590	V=4	<0.0001	robust.recalibr
robust.recalibr.γ=2	min.pivot	-0.0889	-0.1092	V=51	<0.0001	robust.recalibr
robust.recalibr.γ=2	surp.overshoot	-0.0459	-0.0220	V=178	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	know.weight	-0.0465	-0.0327	V=106	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	meta.prob.weight	-0.0679	-0.0675	V=1	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	min.pivot	-0.0998	-0.1152	V=52	<0.0001	robust.recalibr
robust.recalibr.γ=2.5	surp.overshoot	-0.0569	-0.0295	V=146	<0.0001	robust.recalibr
robust.recalibr.γ=3	know.weight	-0.0533	-0.0361	V=99	<0.0001	robust.recalibr
robust.recalibr.γ=3	meta.prob.weight	-0.0748	-0.0740	V=7	<0.0001	robust.recalibr
robust.recalibr.γ=3	min.pivot	-0.1066	-0.1174	V=58	<0.0001	robust.recalibr
robust.recalibr.γ=3	surp.overshoot	-0.0637	-0.0351	V=138	<0.0001	robust.recalibr

Table F4: Comparison of Brier scores, two-sided paired Wilcoxon signed rank tests, robust recalibration vs benchmarks in each data set.