Mach Learn (2014) 95:261–289
DOI 10.1007/s10994-013-5401-4

Forecast aggregation via recalibration

Brandon M. Turner · Mark Steyvers · Edgar C. Merkle · David V. Budescu · Thomas S. Wallsten

Received: 25 July 2012 / Accepted: 25 July 2013 / Published online: 12 October 2013

Abstract It is known that the average of many forecasts about a future event tends to outperform the individual assessments. With the goal of further improving forecast performance, this paper develops and compares a number of models for calibrating and aggregating forecasts that exploit the well-known fact that individuals exhibit systematic biases during judgment and elicitation. All of the models recalibrate judgments or mean judgments via a two-parameter calibration function, and differ in terms of whether (1) the calibration function is applied before or after the averaging, (2) averaging is done in probability or log-odds space, and (3) individual differences are captured via hierarchical modeling. Of the non-hierarchical models, the one that first recalibrates the individual judgments and then averages them in log-odds is the best relative to simple averaging, with 26.7 % improvement in Brier score and better performance on 86 % of the individual problems. The hierarchical version of this model does slightly better in terms of mean Brier score (28.2 %) and slightly worse in terms of individual problems (85 %).

Keywords Calibration · Aggregation · Forecasting · Systematic distortions · Hierarchical Bayesian models · Individual differences · Wisdom of the crowd

Editors: Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach.

B.M. Turner
Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

M. Steyvers (✉)
University of California, Irvine, CA 92697, USA
e-mail: mark.steyvers@uci.edu

E.C. Merkle
University of Missouri, 319 Jesse Hall, Columbia, MO 65211, USA

D.V. Budescu
Fordham University, 441 East Fordham Road, Bronx, NY 10458, USA

T.S. Wallsten
University of Maryland, 8082 Baltimore Avenue, College Park, MD 20740, USA

1 Introduction

In many situations, experts are asked to provide subjective probability or confidence estimates of uncertain events. The estimates can relate to general knowledge questions (e.g., Which city is furthest North of the equator, Rome or New York?) or to predicting event occurrence (e.g., Will political candidate X win the election?). There exists a large body of work focused on the use of statistical models for combining these individual subjective probability judgments into a single probability estimate (e.g., Ariely et al. 2000; Clemen 1986; Clemen and Winkler 1986; Cooke 1991; Wallsten et al. 1997). A simple form of aggregation, namely the unweighted linear average, has proven to be effective in many situations (e.g., Armstrong 2001). Alternatively, the aggregation can be accomplished by taking a weighted average of the reported probability estimates, with, for example, weights determined by previous expert performance (Cooke 1991).

We consider the aggregation of subjective forecasts voluntarily provided by users of a website. This is similar in spirit to the machine learning notion of aggregating across weak learners, as implemented in popular methods such as bagging (Breiman 1996) and boosting (Freund and Schapire 1996). The former method involves fitting a series of weak learners (often classification or regression trees) to bootstrapped samples of the data, with overall predictions obtained by averaging across the weak learners. The latter method involves fitting a series of weak learners to weighted versions of the original dataset, with overall predictions obtained by taking an accuracy-weighted average across the weak learners. The aggregation of weak machine learning algorithms differs from the aggregation of human forecasts in that (i) given a domain, the algorithms often exhibit more stable behavior, and (ii) the humans often contribute only a small number of forecasts. While the ideas of averaging and weighting weak learners definitely translate to the aggregation of human forecasts, specific implementations must deal with these additional data sparsity and variability problems.

An important consideration for aggregation approaches involving human forecasts is the presence of systematic biases that might distort the individuals' subjective probability estimates. For example, when using probabilities to report confidence in one's judgment, individuals often report values that are too extreme (e.g., Brenner et al. 1996; Keren 1991; Lichtenstein et al. 1982; Yates 1990). Merkle (2010) estimated and corrected for these systematic biases in psychological data. Further, Shlomi and Wallsten (2010) have shown that judges are sensitive to miscalibrated subjective probabilities, and they are able to internally recalibrate miscalibrated information from advisors. These findings suggest that we might improve forecast aggregation by correcting for forecasters' systematic biases. In this article, we construct a series of models that first estimate the bias inherent in judges' forecasts, then correct and aggregate the forecasts into a single value. The models we investigate differ in both the extent to which they accommodate individual differences and where the bias correction takes place (see Clemen 1989, for a related discussion of the latter issue).

We first present the general recalibration function used in all the models, and then describe five aggregation models that use the function in different ways. We apply the models to data from a recent forecasting study, comparing the models to one another and to the unweighted average forecast. By investigating a wide variety of model variants, we seek to understand which modeling procedure leads to the most accurate forecasting performance, as measured by a cross validation test. Our research centers on three questions. (1) Is it better to aggregate the raw individual judgments and then recalibrate the average; to first recalibrate the individual judgments and then average those values; or to recalibrate the individuals, average those values and then recalibrate that average? (2) Regardless of the answer to the first question, how should the recalibration take place—on the probabilities themselves or following some transformation, such as log odds? (3) Does including individual differences in these models improve accuracy or does it simply reduce their generalizability to new questions? We investigate each of these questions, draw conclusions about optimal calibration methods, and relate these conclusions to possibilities for future methods and applications.

2 Recalibrating subjective probability estimates

A large body of evidence suggests that subjective probability estimates systematically deviate from objective measures (see Zhang and Maloney 2011, for examples across many research domains). In forecasting situations, the probability of rare events is often overestimated while the probability of common events is underestimated. This tendency is related to the miscalibration that is often found in psychology research. For example, judges consistently overestimate the probability of precipitation (Lichtenstein et al. 1982). Murphy and Winkler (1974) asked judges to first report the probability of precipitation for the next day. After this initial estimate, judges were provided with information from a computerized weather prediction system, and were asked to reestimate their probabilities. The manipulation showed no effect and both responses demonstrated overestimation of the probability of precipitation.

Miscalibration in prediction often carries over to natural or expertise domains, but not always. Griffin and Tversky (1992) found that when judges were asked about attributes such as population size of pairs of states, they produced significantly overconfident responses. In the prediction task, Murphy and Winkler (1984) showed that professional weather forecasters are remarkably well-calibrated, producing nearly perfect probability estimates. However, Christensen-Szalanski and Bushyhead (1981) showed that when physicians were asked to estimate the likelihood of pneumonia in patients, they were grossly overconfident (predicting probabilities as extreme as 0.88 when the actual probability was only about 0.14). This difference can be explained by the fact that weather forecasters make large numbers of forecasts and receive relatively immediate feedback on them, while physicians do not (Wallsten and Budescu 1983; Fryback and Erdman 1979). The forecasting situations that are considered below tend to be more similar to those made by physicians than to those made by weather forecasters.

In other experimental paradigms, Brown and Steyvers (2009) asked judges to both infer and predict stimulus properties in a perceptual task involving four alternatives. Their experiment consisted of an "inference" task in which subjects were instructed to choose the alternative that was most likely to have produced a particular stimulus, and a "prediction" task in which judges were asked to choose the alternative that was most likely to produce the next stimulus (i.e., a prediction about a *future* stimulus). On each trial, one of the four alternatives produced a stimulus in a manner that induced a strong positive sequential dependency. Specifically, the alternative that produced the stimulus on Trial t was more likely than the others to produce the stimulus on Trial $t+1$. By using the same stimulus set in both the inference and prediction tasks, Brown and Steyvers found that 48 of the 63 judges estimated a higher probability of a change in the prediction task than they did in the inference task, suggesting greater miscalibration in the former than in the latter (see also Wright and Wisudha 1982; Wright 1982).

As mentioned earlier, judges who accurately estimate the observed relative frequency of events are said to be "calibrated." That is, if a judge reports a probability of occurrence of p, and the event happens pn out of n times, then the judge is well recalibrated. Calibration can be assessed both visually and statistically. A common empirical approach involves plotting observed relative frequencies (y-axis) conditional on judged or subjective probabilities (x-axis). In this approach, the relative frequency of occurrence is computed by binning the subjective probabilities and computing the proportion of events that occurred within each bin. The line of perfect calibration is then $y = x$, with data falling below (above) the line implying overconfidence (underconfidence). These plots can also be reinforced with simple statistical measures of calibration, often based on decompositions of the Brier score (Arkes et al. 1995; Yates 1982).

An alternative approach involves fitting a function to the (X, Y) plot that characterizes the nature and extent of the deviation of the points from the diagonal. The function itself can be used to "recalibrate" the judged probabilities, while estimates of the function's parameters can be used to compare the extent and nature of miscalibration across judges, groups, or experiments. Many different functions are available, but because of its great flexibility we decided to use the Linear in Log Odds (LLO) function. Our use is different from previous decision researchers, however. Whereas they were concerned with decision weights in choice tasks, our focus is on transforming judged probabilities to render them more accurate.

2.1 The linear in log odds function

The LLO recalibration function has been used extensively as a method for estimating the distortion of subjective individual probability estimates from their true experimental probabilities in the context of risky decision (e.g., Birnbaum and McIntosh 1996; Gonzalez and Wu 1999; Tversky and Fox 1995). To derive the functional form, we assume that the recalibration function $c(p)$ is linear with respect to p on the log odds scale, so that

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

where γ and τ are the slope and intercept, respectively. If we solve for $c(p)$ in Eq. (1), we obtain the recalibration function

$$c(p|\gamma,\delta) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}, \tag{2}$$

where $\delta = \exp(\tau)$. The slope parameter γ in Eq. (1) corresponds to the curvature of the function in Eq. (2) and the intercept parameter $\tau = \log(\delta)$ controls the height above zero.

In analogy to Gonzalez and Wu's (1999) argument about the weighting function parameters, the recalibration function provides a convenient functional form with two psychologically interpretable parameters. The first parameter γ in our use of the model corresponds to discriminability, which manifests itself in the functional form by means of curvature. As γ increases, the form of the calibration function becomes more step-like, indicating that judges' estimates of low and high probability events are insufficiently extreme. The second parameter δ represents overall response tendency, which is represented as the vertical distance of the curve from zero. Tendencies for higher estimates yield higher calibration curves.

Figure 1 shows various LLO curves as a function of different parameter values. The left panel shows how the parameter γ affects the functional form by fixing $\delta = 0.6$ and incrementing γ by 0.1 from $\gamma = 0$ to $\gamma = 3$. The figure (left panel) shows that as γ is increased from zero to three, the curves go from being a straight horizontal line to sharply increasing on the interval [0.3, 0.8]. The right panel shows how δ affects the functional form by fixing $\gamma = 0.2$ and incrementing δ by 0.1 from $\delta = 0$ to $\delta = 3$. The figure (right panel) shows that as δ is increased from zero to three, the function increases in height above zero (also known as "elevation"; Gonzalez and Wu 1999).

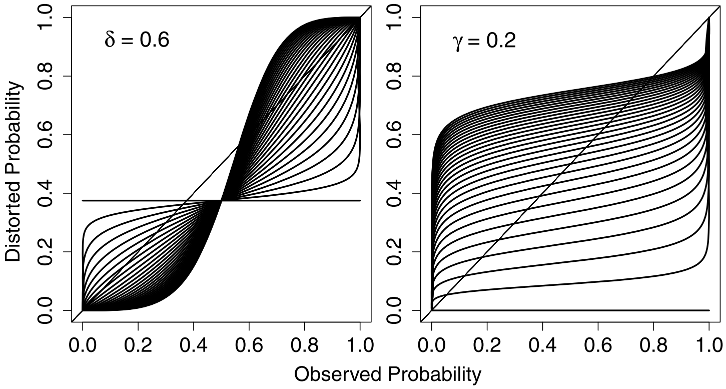

Fig. 1 Various LLO curves under different parameter values. The *left panel* shows how the γ parameter affects the functional form by fixing $\delta = 0.6$ whereas the *right panel* shows how δ affects the functional form by fixing $\gamma = 0.2$

Figure 1 shows that the function is very flexible and has a number of interesting properties. First, we notice that under the restriction $\gamma = 1$ and $\delta = 1$, $c(p) = p$ and no transformation is applied. Second, if $\delta = 1$ the function reduces to another well-known calibration curve known as the Karmarkar's equation (Karmarker 1978). Third, the function must go through the points $(0, 0)$ and $(1, 1)$, which is not always true for some calibration functions. Finally, the function is guaranteed to cross the identity line $c(p) = p$ at exactly one location p^* (except when $\gamma = \delta = 1$) where

$$p^* = \frac{\delta^{1/(1-\gamma)}}{1 + \delta^{1/(1-\gamma)}}. \tag{3}$$

Note that the LLO function does not require that the curve passes through the point $(0.5, 0.5)$, nor that it is symmetric. This property implies that the recalibration function is not a probability function, which by definition must satisfy

$$c(p) + c(1 - p) = 1.$$

Functions that do not satisfy this constraint typically provide good fits to empirical data, where the general finding is that judges both overestimate and overweight low-probability events and even more dramatically underestimate and underweight high-probability events (e.g., Camerer and Ho 1994; Tversky and Kahneman 1992; Wu and Gonzalez 1996).

Once a recalibration function has been fit to data for a particular judge, if the sum of the recalibrated values for complementary probabilities is less than one (i.e., $c(p) + c(1 - p) < 1$), the judge is said to exhibit *subcertainty* (Kahneman and Tversky 1979). In our applications, the very flexible form of the LLO recalibration function will be advantageous because judges appear to be overly optimistic about the probability of future outcomes (i.e., they tend to provide judgments that are higher than the relative frequency of an event occurring; see Brown and Steyvers 2009).

To estimate the parameters γ and δ in Eq. (2), one can use classic maximum likelihood methods. If we let X_j indicate whether the jth event did ($X_j = 1$) or did not occur ($X_j = 0$), then the likelihood function is the product

$$L(\gamma, \delta|X) = \prod_i \left\{ c(p|\gamma, \delta)^{X_j} \left[1 - c(p|\gamma, \delta)\right]^{1-X_j} \right\}, \tag{4}$$

where p is a single forecast associated with Event j. To obtain the maximum likelihood estimates, one uses standard numerical optimization routines to optimize Eq. (4) with respect to γ and δ.

If there are enough data, an alternative to likelihood-based methods are nonparametric estimation techniques (Gonzalez and Wu 1999; Page and Clemen 2012). For example, Gonzalez and Wu proposed a nonparametric estimation algorithm that returns estimates of the value and recalibration function in the prospect theory model (Kahneman and Tversky 1979), but other recalibrating functional forms can be estimated using this approach. Gonzalez and Wu found that using the nonparametric approach provided a great deal of flexibility in fitting the data from a calibration experiment. As another example, Page and Clemen used a localized kernel density estimator (see Silverman 1986) in combination with a clustered bootstrap approach (see Härdle 1992) to fit calibration curves to data from a prediction market.

One can also employ Bayesian methods to estimate the parameters γ and δ in Eq. (2). In the Bayesian framework, one assumes that the parameters, along with the data, are random quantities (e.g., Gelman et al. 2004). In contrast to classical statistics, inference about the parameters are based on their probability distributions after some data are observed (see Christensen et al. 2011; Gelman et al. 2004). We rely on Bayesian estimation procedures in the applications described below.

2.2 Linear in log odds aggregation

For the longitudinal dataset to which we apply the models, some events do not yet have known outcomes. We incorporate these missing observations into the vector x_j, the result of the jth event. When $x_j = 1$, the event did occur—an event we refer to as a *resolved* event—whereas $x_j = 0$ denotes that the event did not occur—an event we refer to as an *unresolved* event. We let $y_{i,j}$ represent the probability estimate provided by Judge i on Event j. Because Eq. (1) is not defined when $p = 0$ or $p = 1$, we must use a correction to judgments such that $y_{i,j} = 0$ or $y_{i,j} = 1$ prior to use in the models. Thus, we adjust these boundary forecasts to 0.001 and 0.999, respectively, prior to fitting any of the models to facilitate a direct evaluation across models.

We generally compare our re-calibration models to the unweighted linear average (sometimes called the *Unweighted Linear Opinion Pool*; ULinOP) of the estimates provided by each of the judges. Thus, predictions $\widehat{\mu}_j$ for ULinOP are obtained by evaluating

$$\widehat{\mu}_j = \frac{1}{n}\left(\sum_{i=1}^{n} y_{i,j}\right),$$

where n is the number of responses obtained on Event j.

Despite its simplicity, the ULinOP is a formidable estimate for the probability of future outcomes in forecasting future events. Some authors have even argued that it is difficult to beat the ULinOP by more than 20 % (e.g., Armstrong 2001).

The models described below each allow for distortions that yield non-additive probabilities. Although we assume that this distortion is a consequence of the functional form in Eq. (2), we exploit the recalibration function in a variety of ways. All of the models presented below can be represented in the general form

$$\text{Model}(y) = f\big(g(y)\big), \tag{5}$$

Table 1 Model specification as a function of an inner function $g(\cdot)$ and an outer function $f(\cdot)$ according to Eq. (5)

Model	Inner $g(\cdot)$	Outer $f(\cdot)$
ULinOP	y_i	$(1/n)\sum_{i=1}^{n} g(y_i)$
Average then Recalibrate	$(1/n)\sum_{i=1}^{n} y_i$	$c(g(y)\vert\gamma,\delta)$
Calibrate then Average	$c(y_i\vert\gamma,\delta)$	$(1/n)\sum_{i=1}^{n} g(y_i)$
Calibrate then Average Log Odds	$\log(\frac{c(y_i\vert\gamma,\delta)}{1-c(y_i\vert\gamma,\delta)})$	$\frac{\exp[(1/n)\sum_{i=1}^{n} g(y_i)]}{1+\exp[(1/n)\sum_{i=1}^{n} g(y_i)]}$

where y denotes the set of observed responses, Model(y) represents the predictions of the model, and $g(\cdot)$ and $f(\cdot)$ are two functions that are either calibration or aggregation functions, depending on the model. Table 1 shows the functions $g(\cdot)$ and $f(\cdot)$ for each model under consideration. The first type of model we present, *Average then Recalibrate*, first averages all responses for Event j and then calibrates the average to estimate the probability of an event occurring. The second type of model recalibrates each individual judgment using the parameters γ and δ, and then averages these recalibrated judgments to estimate the event probability. We explore two variants of this model. In the first version, *Calibrate then Average*, the averaging is performed directly on the recalibrated judgments. The second variant of this model, *Calibrate then Average Log Odds*, performs the averaging on the log odds of the recalibrated judgments. We will show that this second variant leads to much better aggregation results. Finally, we also examine hierarchical extensions of the recalibration model (not shown in Table 1) that incorporate individual differences into the estimation of the parameters γ and δ in Eq. (2).

2.3 Average then recalibrate model

The first model we consider averages the responses for a particular event and then recalibrates the average. One useful way to view this model is as a transformation of the ULinOP discussed previously. Instead of taking the group average at face value, the group average is transformed using Eq. (2). While this type of modeling does not necessarily have clear psychological interpretability (as individual differences are ignored), it dampens the impact of extreme predictions (i.e., zeros or ones) given by individual judges. In addition, averaging the individuals' biases may produce more stability in the estimation of the calibration parameters.

Thus, for the jth resolved event, we assumed that

$$p_j = \frac{1}{S_j}\sum_{i\in\mathcal{Q}_j} y_{i,j},$$

$$\mu_j = c(p_j|\gamma,\delta), \quad \text{and}$$

$$x_j \sim \text{Bernoulli}(\mu_j),$$

where p_j is the average probability elicited by judges for the jth event, x_j is the coded known outcome for the jth event, $\mathcal{Q}_j$ is the set of judges who responded to the jth event, S_j is the number of judges who responded to the jth event (i.e., $S_j = |\mathcal{Q}_j|$), and $c(\cdot|\gamma,\delta)$ is governed by Eq. (2). Thus, for the jth event, the likelihood function can be written as

$$L(\gamma, \delta | x_j, y_{i,j}) = \left[c\left(\frac{1}{S_j} \sum_{i \in Q_j} y_{i,j} \mid \gamma, \delta \right) \right]^{x_j} \left[1 - c\left(\frac{1}{S_j} \sum_{i \in Q_j} y_{i,j} \mid \gamma, \delta \right) \right]^{1-x_j}. \tag{6}$$

As stated previously, we estimated the models via Bayesian methods that require prior distributions. After some inspection, we settled on mildly informative priors for each of the model parameters here so that

$$\begin{aligned} \delta &\sim \Gamma(1, 1) \quad \text{and} \\ \gamma &\sim \Gamma(1, 1), \end{aligned} \tag{7}$$

where $\Gamma(a, b)$ denotes the gamma distribution with rate a and shape parameter b. The $\Gamma(1, 1)$ prior has a mean and standard deviation of 1, and a 95 % credible set of approximately (0.025, 3.703). We chose these priors after observing the shape of the calibration function for a representative range of values for γ and δ within this range. We note that a $\Gamma(1, 1)$ prior is equivalent to an exponential prior with rate parameter equal to one (i.e., $\Gamma(\alpha, 1) = \text{Exp}(\alpha)$ for some rate parameter α).

With our fully-specified model, we can now write the joint posterior distribution for γ and δ as

$$\pi(\gamma, \delta | x, y) \propto \prod_{j=1}^{J} L(\gamma, \delta | x_j, y_{i,j}) \pi(\gamma) \pi(\delta), \tag{8}$$

where $L(\gamma, \delta | x_j, y_{i,j})$ is defined in Eq. (6), and J is the number of resolved events.

Figure 2 shows a graphical diagram for this model. These types of diagrams are often very useful for illustrating how the parameters in the model (white nodes) are connected via arrows to the observed data (gray nodes; see Buntine 1994; Lee 2008; Lee and Wagenmakers 2012; Shiffrin et al. 2008). When the variables are discrete-valued, they are shown as square nodes, whereas when the variables are continuous, they are shown as circular nodes. A double bordered variable indicates that the quantity is deterministic, not stochastic. Finally, "plates" show how vector-valued variables are interconnected. For example, in Fig. 3, we see that the nodes γ and δ are not on the plate, which indicates that these parameters are fixed across events, whereas there are separate μ_js for each Event j ranging from one to J, which are connected to the other nodes on the plate.

To elicit a prediction for Event j, the model first calculates the mean of the posterior distribution for each of the parameters μ_j, so that

$$\widehat{\mu}_j = \frac{1}{K} \left(\sum_{k=1}^{K} \mu_{j,k} \right), \tag{9}$$

where K is the number of samples drawn from the posterior distribution (see below for more details), and $\mu_{j,k}$ is the kth sample of the posterior corresponding to μ_j. Once each of these μ_js are obtained, the model returns the estimate

$$\{\widehat{\mu}_j, 1 - \widehat{\mu}_j\} \tag{10}$$

for the probability of the event occurring and the probability of the event not occurring, respectively.

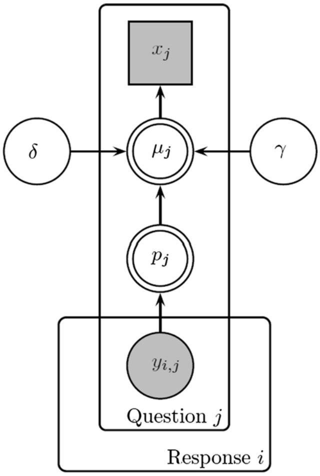

Fig. 2 Graphical diagram of the Average then Recalibrate Model

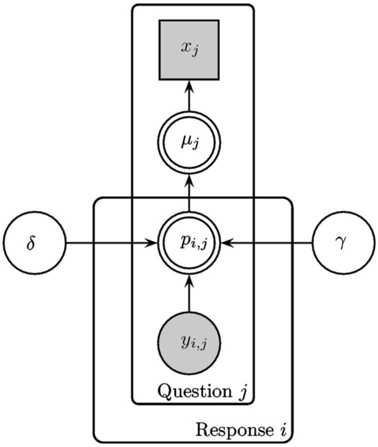

Fig. 3 Graphical diagram of the Recalibrate then Average Model

2.4 Calibrate then average model

The next calibration model we examined first recalibrates the reported probabilities for each judge and then averages the results across judges to produce a single group forecast. Specifically, this model applies the calibration function shown in Eq. (2) to the observed responses $y_{i,j}$ (under the assumption that parameters are equal across subjects), creating the auxiliary variable $p_{i,j}$. For resolved events, we assume that

$$\begin{aligned} p_{i,j} &= c(y_{i,j}|\gamma, \delta), \\ \mu_j &= \frac{1}{S_j} \sum_{i \in \mathcal{Q}_j} p_{i,j}, \quad \text{and} \\ x_j &\sim \text{Bernoulli}(\mu_j). \end{aligned} \tag{11}$$

Thus, for the jth event, the likelihood function can be written as

$$L(\gamma, \delta | x_j, y_{i,j}) = \left[\frac{1}{S_j} \sum_{i \in \mathcal{Q}_j} c(y_{i,j} | \gamma, \delta) \right]^{x_j} \left[1 - \frac{1}{S_j} \sum_{i \in \mathcal{Q}_j} c(y_{i,j} | \gamma, \delta) \right]^{1-x_j}. \tag{12}$$

For this model, we again assumed informative priors shown in Eq. (7). Thus, the joint posterior distribution for γ and δ is as specified in Eq. (8), where $L(\gamma, \delta | x_j, y_{i,j})$ is now given by Eq. (12). Figure 3 shows a graphical diagram for this model. To make a prediction, the model forms an estimate by calculating Eq. (9) and then returning the estimate as in Eq. (10).

2.5 Calibrate then average on the log odds scale

For the Recalibrate then Average models, we investigated two different methods of aggregating the recalibrated individual judgments. The first method, discussed above, averages the recalibrated judgments on the probability scale. A problem with averaging a set of recalibrated judgments is that the average may not necessarily produce an optimally calibrated model prediction (Hora 2004). The problem occurs in the transition from the recalibrated judgments $p_{i,j}$ to the aggregated model prediction μ_j. For a given distribution of elicited judgments, the application of Eq. (2) (i.e., when $\gamma \neq \delta \neq 1$) results in model predictions that are uncalibrated with respect to the event outcome.[1]

To illustrate the problem, the left panel of Fig. 4 shows a distribution of individual probability judgments, represented as y (bottom histogram). After recalibrating these judgments via a LLO recalibration curve with $\gamma = 2$ and $\delta = 0.5$, the resulting distribution is represented as p (the far left histogram).

The means of the uncalibrated judgments is represented as the vertical dashed line and the mean of the recalibrated judgments is represented as the horizontal dashed line. If the average of p was a fully recalibrated (i.e., with respect to the event outcome) version of y, the horizontal line would intersect with the vertical line at a point directly on the recalibration curve. The figure shows that although the difference is slight, the average of p is not a recalibrated version of y. Therefore, by first recalibrating individual judgments and then averaging the resulting recalibrated judgments, the average may not be recalibrated.[2]

To remedy the problem of uncalibrated aggregate predictions, we investigated a method of aggregation on the log odds scale, where the LLO calibration function becomes linear. For a given judgment $y_{i,j}$, we first recalibrate the judgment on the log odds scale, so that

$$p_{i,j} = \gamma \log\left(\frac{y_{i,j}}{1 - y_{i,j}} \right) + \log(\delta). \tag{13}$$

This transformation converts the judgments $y_{i,j}$ to recalibrated judgments $p_{i,j} \in (-\infty, \infty)$. We now have recalibrated judgments that are linear with respect to $\log[y_{i,j}/(1 - y_{i,j})]$. We can now aggregate these judgments in the log odds scale

$$\mu_j^* = \frac{1}{S_j} \sum_{i \in \mathcal{Q}_j} p_{i,j}.$$

[1] Producing an uncalibrated model prediction by aggregating calibrated forecasts is an example of Jensen's inequality.

[2] The judgments are always fully recalibrated in the trivial case when $\gamma = \delta = 1$.

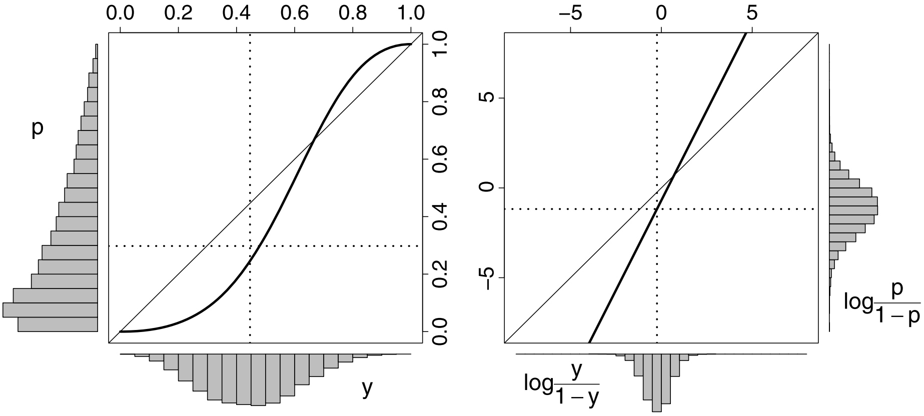

Fig. 4 A comparison of calibration distortions on the probability scale (*left panel*) and on the log odds scale (*right panel*). The *bottom left histogram* represents a distribution of probability judgments y, and the far *left histogram* represents these judgments after the recalibration. The *bottom right histogram* represents these same judgments on the log odds scale, and the far *right histogram* represents these judgments after recalibration, on the log odds scale

Thus, the quantity μ_j^* can be viewed as a calibrated version of the log of the geometric mean of the odds ratio $y/(1-y)$.[3] The aggregate μ_j^* can then be converted back to the probability scale, producing

$$\mu_j = \frac{\exp(\mu_j^*)}{1+\exp(\mu_j^*)}.$$

As before, we assume that the resolution occurs randomly with probability μ_j, so that

$$x_j \sim \text{Bernoulli}(\mu_j),$$

and assumed the mildly informative priors shown in Eq. (7).

The right panel of Fig. 4 shows this aggregation scheme on the log odd scale. We begin with a distribution of individual probability judgments converted to log odds, which we represent as $\log(y/(1-y))$ (bottom histogram). This distribution is then recalibrated by evaluating Eq. (13), creating the distribution of recalibrated judgments on the log odd space, which we denote $\log(p/(1-p))$ (right histogram). As indicated by the dashed vertical and horizontal lines, the mean of the recalibrated individual judgments, $\log(p/(1-p))$, is recalibrated, so once it is converted to probability space, the aggregate can be naturally used to model the Bernoulli outcome x_j.

2.6 Hierarchical recalibrate then average model

We also examined a hierarchical extension of the Recalibrate then Average Model presented above. As with the previous model, we again recalibrate the reported probabilities from individual forecasters and then average the results. However, instead of assuming a single set of calibration parameters across all individuals, we now assume that each judge i has

[3] Furthermore, when $\delta = \gamma = 1$, no calibration occurs and μ_j^* equals the log of the geometric mean of the odds ratio.

a different calibration function associated with her own parameter set, thereby allowing us to capture individual differences. In the context of calibration, hierarchical models have been shown to drastically improve the interpretation and precision of inferential analyses in experimental studies (e.g., Budescu and Johnson 2011; Merkle et al. 2011).

Similar to the Calibrate then Average model above, for the Hierarchical Recalibrate then Average model, we first recalibrate each judge's response $y_{i,j}$ through the LLO function (see Eq. (2)), so that

$$p_{i,j} = c(y_{i,j}|\gamma_i, \delta_i),$$

where the parameters γ_i and δ_i are the recalibration parameters for the ith judge. Note that this is different from the Calibrate then Average model, where we assumed a single γ and δ across all judges. As a result, we cannot simply aggregate the $p_{i,j}$s and connect the aggregate to the event resolution vector x_j (see Eq. (11)) because we will be unable to estimate each individual judge's calibration parameters. Because of the dimension mismatch between the matrices p and x, to connect these matrices in the model we must define an auxiliary matrix x^* whose individual elements $x^*_{i,j}$ contain the event resolution information for the ith judge on the jth item. Here we make a distinction between the generative process and the inferential process. By creating the auxiliary matrix x^*, the generative process would allow for different event resolution information for each individual judge. This feature of the model would be useful if we were interested in examining the effects of accurate feedback on the calibration parameters. However, in this article we assume in the inferential process that the event resolution is the same for each individual judge such that

$$x^*_{1,j} = x^*_{2,j} = \cdots = x^*_{S_j,j} = x_j$$

for all S_j judges who responded to Item j. We assume that the answer to the jth resolved event (for the ith judge) is a Bernoulli random variable distributed with probability equal to $p_{i,j}$, or

$$x^*_{i,j} \sim \text{Bernoulli}(p_{i,j}).$$

Defining the model in this way allows the calibrated judgments $p_{i,j}$ to vary from one judge to another while holding the event resolutions to be the same across judges. In other words, $p_{i,j}$ is the model's estimate of Judge i's probability that Event j will occur, and therefore $p_{i,j}$ models the Bernoulli process for that judge and item. We assume the calibration parameters are commonly distributed according to one hyper-distribution, so that

$$\gamma_i \sim \Gamma(\gamma_\alpha, \gamma_\beta), \quad \text{and}$$
$$\delta_i \sim \Gamma(\delta_\alpha, \delta_\beta).$$

After some inspection, we arrived at the following mildly informative priors for the hyperparameters:

$$\delta_\alpha \sim \Gamma(1000, 1000),$$
$$\gamma_\alpha \sim \Gamma(1000, 1000),$$
$$\delta_\beta \sim \Gamma(1000, 1000) \quad \text{and}$$
$$\gamma_\beta \sim \Gamma(1000, 1000).$$

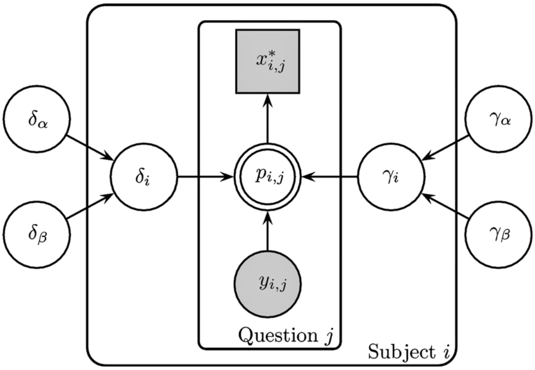

Fig. 5 Graphical diagram of the Hierarchical Calibration Model

We chose these priors in part to maintain consistency among the models. The $\Gamma(1000, 1000)$ prior has a mean of 1, a standard deviation of 0.032, and a 95 % credible set of approximately (0.939, 1.063). Thus, when there are very few observations, the hyperparameters $(\delta_\alpha, \gamma_\alpha, \delta_\beta, \gamma_\beta)$ become approximately equal to one, nearly translating to the priors

$$\gamma_i \sim \Gamma(1, 1), \quad \text{and}$$
$$\delta_i \sim \Gamma(1, 1),$$

as specified in the previous models.

Once we have established an estimate for each $p_{i,j}$, we aggregate the estimate across the judges to produce a prediction μ_j from the model, so that

$$\mu_j = \frac{1}{S_j} \sum_{i \in \mathcal{Q}_j} p_{i,j}.$$

Figure 5 shows a graphical diagram for the Hierarchical Recalibrate then Average model. Unlike the previous models, μ_j is not part of the generative process, rather it is part of the inference process. As a result, μ_j does not appear in Fig. 5, although it is used to elicit a prediction from the model.

2.7 Hierarchical recalibrate then average on the log odds scale

We also implemented a hierarchical version of the Recalibrate then Average on the Log Odds Scale model. We used the same priors as in the Hierarchical Recalibrate then Average model for all parameters and we applied the same transformation that we used in the Recalibrate then Average Log Odds model (see, for example, Eq. (13)). Although the outputs from this model are produced from an aggregate on the log odds scale, the predictions may be somewhat miscalibrated because of the influence of the prior and the sparsity in the individual judgments.

We now present the results of fitting the models to the data. After fitting each model to our data set, we compare each one to the ULinOP and then to one another via their Brier scores. We further illustrate the differences between these models by presenting the estimated posterior distributions of the parameters and the posterior predictive distributions of the models.

3 The data

To test the models' aggregation abilities, we use data collected by the Aggregative Contingent Estimation System (ACES), a large-scale project for collecting and combining forecasts of many widely-dispersed individuals (http://www.forecastingace.com/aces). A preliminary description of the data collection procedure can be found in Warnaar et al. (2012). Volunteer participants were asked to estimate the probability of various future events' occurrences, such as the outcome of presidential elections in Taiwan and the potential of a downgrade of Greek sovereign debt. Participants were free to log on to the website at their convenience and forecast any items of interest. The forecasting problems involved events with two outcomes (event X will occur or not) as well as more than two outcomes (e.g., event A, B, or C will occur). For this paper, we focused on a subset of 176 resolved forecasting problems that met a number of constraints. First, they involved binary events only because we are only considering calibration models for binary events. Second, all 176 forecasting problems involved a standard way of framing the event and was presented in the form: will event X happen by date Y? This last constraint excluded a small number of events where the event was framed in terms of a deviation from status quo (e.g. will X remain true by date Y?). Finally, we only included forecasting problems where the event of interest could happen at any time before the deadline associated with the event. For example, in the event "Greece will default on its debt in July 2011", the target event could have occurred anytime before the end of July (it did not). However, items such as "The Cowboys and Aliens comic book movie will out-gross the Green Lantern movie on its opening weekend July 29th", were not included because the event cannot happen on any other day except July 29th. These latter items were excluded as a result of their framing because their status quo could not be altered. A total of 1401 participants contributed judgments to these forecasting problems.

3.1 Model scoring through cross-validation

We primarily evaluate models through use of the Brier score (Armstrong 2001; Brier 1950; Murphy 1973). The scoring rules that are commonly used in forecasting are often called loss functions in the machine learning literature (e.g., Hastie et al. 2009). Popular loss functions for binary forecasts include squared-error loss and the (negative of the) Bernoulli log-likelihood, which forecasting researchers sometimes call the "Brier score" and "logarithmic score," respectively. These loss functions have the desirable property that they are "strictly proper" (see, e.g., O'Hagan et al. 2006), meaning that the forecaster minimizes her expected loss only by reporting her true beliefs. The expected loss cannot be reduced under alternative strategies, such as reporting forecasts that are more extreme than one's true beliefs. The above loss functions have also been shown to generally yield similar conclusions in a forecasting context (e.g., Staël von Holstein 1970), leading us to focus on the Brier score (squared error loss) in this paper.[4]

Because all events involved only two outcomes, the Brier score for the jth event can be expressed as

$$B_j = (X_j - \widehat{\mu}_j)^2,$$

[4] We also evaluated the models with both spherical and logarithmic scoring rules. However, because the results were invariant to these scoring rules (also see Staël von Holstein 1970), we present only the results using the Brier score.

where $\widehat{\mu}_j$ is the model prediction for event j's occurrence and X_j is the resolution of the jth event. For example, if the jth event did occur, then $X_j = 1$. Thus, in this definition of the Brier score, the best score B_j is zero, and the worst possible score is one.

We calculate the Brier scores in a 10-fold cross-validation procedure where the parameters of the calibration models are estimated on a subset of 90 % of the forecasting problems. For the remaining 10 % of forecasting problems, the resolution is withheld from the model, and the calibration models are used to make the predictions $\widehat{\mu}_j$ for only that subset of forecasting problems. In the 10-fold cross-validation procedure, this process of estimation and prediction is repeated on 10 random and non-overlapping partitions of the data set to create a full set of model predictions $\widehat{\mu}_j$.

After the out-of-sample Brier scores are obtained for each event, we compute the mean predictive error (MPE) by averaging the Brier scores across the number of events J, so

$$\text{MPE} = \frac{1}{J}\sum_{j=1}^{J} B_j.$$

Once the MPE scores are calculated for each model, we evaluate the percentage improvement over a baseline model, which in our case is the unweighted linear average (ULinOP). We calculate the percentage difference of the mean (PDM) in MPE for the ith model, denoted MPE(i) relative to the score for the ULinOP MPE(0), by calculating

$$\text{PDM} = 100 \times \frac{\text{MPE}(0) - \text{MPE}(i)}{\text{MPE}(0)}. \tag{14}$$

Therefore, PDM values larger than 0 indicate that the calibration model improves over the average prediction of the unweighted linear average (ULinOP).

In addition to the global measures of model performance given by MPE and PDM, we will also compare models at the individual event level using a pair-wise procedure proposed by Broomell et al. (2011). In this procedure, we calculate PW(a, b), the number of events for which model a has a lower Brier score than model b

$$\text{PW}(a, b) = \sum_{j=1}^{J} \mathbb{1}_{B_{a,j} < B_{b,j}} \tag{15}$$

where $B_{a,j}$ is the Brier score for the jth event for model a, and $\mathbb{1}_x$ is an indicator function. This pair-wise comparison is useful because one model might have a higher MPE relative to another model but still be the better model in the pair-wise comparison. Therefore, the MPE score measures how well the model performs on average, whereas the PW score measures relative model performance on an individual event basis.

To fit all the models, we used the program JAGS (Plummer 2003) to estimate the joint posterior distribution of each set of model parameters (code for the models appears in the appendices). For each model, we obtained 1,000 samples from the joint posterior after a burn-in period of 1,000 samples, and we also collapsed across two chains. For the Recalibrate then Average and Average then Recalibrate models we initialized each of the chains by setting $\gamma = 0.17$ and $\delta = 0.34$. For the Hierarchical Calibration model, we initialized each of the chains by setting $\gamma_\alpha = \delta_\alpha = 3$ and $\gamma_\beta = \delta_\beta = 6$.

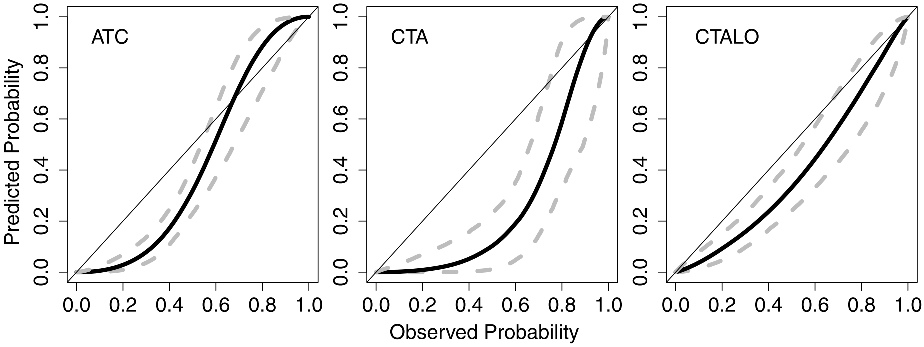

Fig. 6 Posterior predictive distributions for each of the three single-level models: Average then Recalibrate (*left panel*), Recalibrate then Average (*middle panel*), and the Recalibrate then Average Log Odds (*right panel*). The median and 95 % credible set for the posterior predictive distributions are shown as the *solid black* and *dashed gray lines*, respectively

4 Results

Table 2 shows a subset of 40 representative forecasting problems. For each problem, a short description of the event is given, as well as the number of judges (S_j), the number of days that the event was active (Days), and the resolution of the event (X_j). Table 3 summarizes model performance with the average Brier score (MPE), and the percentage improvement over the baseline model (PDM). The table also shows the 95 % confidence interval for the MPE and PDM, which we obtained by a bootstrapping procedure. The bootstrapping was performed by repeatedly sampling forecasting problems and calculating performance statistics on subsets of the forecasting problems in a 10-fold cross validation analysis. In the sections below, we first describe the performance of each individual model in detail and then discuss the pair-wise comparison scores (PW) between models (Table 4).

4.1 Average then recalibrate model

To evaluate the model, we first examined the posterior distribution of the parameters and the posterior predictive distributions of the model. Table 5 summarizes the marginal distributions of γ and δ by providing the median and 95 % Bayesian credible set. The estimates show that there is still a good deal of uncertainty about these parameters, especially given the amount of data. The left panel of Fig. 6 shows the posterior predictive distribution. To plot this distribution, we took 1,000 samples from the estimated joint posterior distribution and produced a calibration function using Eq. (2). We then plotted the median (black line) and 95 % credible set (gray lines).

Note that the model predictions of the ULinOP and the Average then Recalibrate model are connected directly through the LLO function (see Eq. (2)). Therefore, the observed probability in the left panel of Fig. 6 corresponds to the ULinOP prediction and demonstrates that the distortion pattern for probability judgments that was found at the level of individual judges was also found at the aggregate level (i.e., overestimation of unlikely future events and underestimation of likely future events). The posterior median of the intersection point $\widehat{p}^* = 0.679$ (from Eq. (3)), so that ULinOP estimates below 0.7 are mapped to smaller values and ULinOP estimates above 0.7 are mapped to larger values.

Table 3 shows that the MPE for the Average then Recalibrate Model is 0.12, which is 21.9 % better than ULinOP. The far right column of Table 3 shows that the 95 % confidence

Table 2 Descriptions of events and summary statistics for a subset of 40 items. Note: S_j = number of forecasters that responded to the event, Days = number of Days that the event was active, X_j = resolution (coded event did (1) or did not (0) happen)

Id	Description	S_j	Days	X_j
3	Greek debt default	343	18	0
4	US credit rating	437	18	0
6	Libya's Leadership	448	49	1
7	Military coup in Venzuela	369	171	0
8	Troops to Mexico	384	18	0
10	Iran nuclear facility	419	171	0
11	US-Korea trade deal	137	18	0
12	Flu pandemic	315	171	0
15	Earthquake in Japan	206	18	0
16	South-Atlantic Hurricanes	186	171	0
26	Karzai associate resigns	60	13	0
30	Military Organization	50	10	0
31	US-Russian Missile Agreement	61	40	0
32	Newt Gingrich's campaign	50	10	0
34	CO_2 emission levels	128	163	1
68	Troop Deployments to Congo	130	159	0
70	Future anti-government rebellions	103	37	0
71	Stem-cell funding issue	125	159	0
72	Space program	112	34	0
74	Facebook public offering	95	30	0
75	New York Times shut down	84	26	0
79	N.Korean prisoner release	34	30	0
80	Sr. Military Leadership Misconduct	131	133	1
82	TSA security practices improve	118	57	0
83	Military acquisition cuts	106	85	0
84	BRAC review of military installations	97	144	0
85	Ethnic clashes in China	104	60	0
86	College tuition increases	152	152	0
91	PC tablets vs. iPads	283	122	0
92	Earth's surface temperature	97	27	0
94	2011 Nobel Prize winners	19	68	0
95	3D printing availability	177	114	0
96	McDonalds at DisneyWorld	138	152	0
99	Nook/Kindle Sales comparison	200	145	0
100	UK Royal Heir	125	145	0
103	Six-Party talks	105	124	0
105	EU candidacy of Serbia	110	124	0
108	Italian debt default	169	124	0
110	WTO membership for Russia	123	108	1
...	...	...	...	...
Mean over 176 events:	–	75.3	54.7	0.210

Table 3 Average prediction error (MPE) and percentage difference of the mean prediction error (PDM) relative to the unweighted average (ULinOP). The ranges provide bootstrap estimates of the 95 % confidence interval

Model	MPE	Bootstrap MPE	PDM	Bootstrap PDM
ULinOP	0.153	(0.136–0.171)	0.0 %	(0.0 %–0.0 %)
Average then Recalibrate	0.120	(0.091–0.151)	21.9 %	(8.8 %–35.4 %)
Calibrate then Average	0.123	(0.096–0.153)	19.7 %	(5.8 %–33.5 %)
Hier. Recalibrate then Average	0.141	(0.111–0.173)	8.0 %	(−9.5 %–24.9 %)
Calibrate then Average Log Odds	0.112	(0.083–0.145)	26.7 %	(13.0 %–41.0 %)
Hier. Recalibrate then Average Log Odds	0.110	(0.083–0.140)	28.2 %	(16.0 %–41.4 %)

Table 4 Pairwise model comparison score PW(a, b): Number of events (%) where the row model (a) has a smaller prediction error than the column model (b). *Key*: UW is ULinOP, ATC is Average then Recalibrate, CTA is Recalibrate then Average, HCTA is Hierarchical Recalibrate then Average, CTALO is Recalibrate then Average Log Odds, HCTALO is Hierarchical Recalibrate then Average Log Odds

Model	UW	ATC	CTA	HCTA	CTALO	HCTALO
UW	–	30 (17 %)	37 (21 %)	40 (23 %)	25 (14 %)	26 (15 %)
ATC	146 (83 %)	–	132 (75 %)	130 (74 %)	66 (38 %)	99 (56 %)
CTA	139 (79 %)	44 (25 %)	–	121 (69 %)	44 (25 %)	57 (32 %)
HCTA	136 (77 %)	46 (26 %)	55 (31 %)	–	41 (23 %)	41 (23 %)
CTALO	151 (86 %)	110 (63 %)	132 (75 %)	135 (77 %)	–	131 (74 %)
HCTALO	150 (85 %)	77 (44 %)	119 (68 %)	135 (77 %)	45 (26 %)	–

Table 5 Summaries of the estimated posterior distributions obtained for all models except the hierarchical calibration models. *Key*: ATC is Average then Recalibrate, CTA is Recalibrate then Average, CTALO is Recalibrate then Average Log Odds

Model	$\widehat{\gamma}$		$\widehat{\delta}$	
	Median	95 % CI	Median	95 % CI
ATC	2.006	(1.301, 2.766)	0.465	(0.305, 0.717)
CTA	1.859	(0.863, 4.025)	0.109	(0.012, 0.300)
CTALO	1.158	(0.811, 1.589)	0.497	(0.317, 0.785)

interval for the bootstrapped percentage improvement is (8.8 %–35.4 %). Because this interval does not contain 0, we can conclude that this model provides a reliable improvement over the basic ULinOP model.

4.2 Calibrate then average

Table 5 summarizes the marginal posterior distributions of γ and δ. The posterior median values were $\widehat{\delta} = 0.109$ and $\widehat{\gamma} = 1.859$, which lead to a posterior median for the intersection point $\widehat{p}^* = 0.915$. The posterior again shows that there is still some uncertainty in the estimates, especially given the amount of data. The bottom left panel of Fig. 6 shows the posterior predictive distribution for this model. Note that the observed probability on the horizontal axis corresponds to the probability judgments from individual judges. The plot again shows a slowly-graded curve indicating that individual judges dramatically overestimated the probability of future events but showed very little underestimation.

The bottom right panel of Fig. 6 shows the posterior predictive distribution for the Recalibrate then Average Log Odds model. It is apparent that the calibration curves for the CTA and the CTALO models are very different. For the log odds version, the median of the marginal posterior distributions (reported in Table 5) are $\widehat{\delta} = 0.497$ and $\widehat{\gamma} = 1.158$, which leads to a posterior median for the intersection point $\widehat{p}^* = 0.915$. Comparing the two calibration functions in the bottom panel of Fig. 6, we see that the probability scale version applies a stronger correction for intermediate probability judgments, pushing them downward toward zero.

The bottom left panel of Fig. 6 emphasizes the need to consider calibration functions that are not probability functions. The fits to this data clearly suggest that the judges in the experiment are subcertain. That is, judges are overly confident in their responses relative to observed outcome frequencies. Furthermore, judges are miscalibrated because the credible intervals of the posterior predictive distributions do not capture the point $(0.5, 0.5)$. Undoubtedly, a large reason for the subcertainty is that for most resolved events, the event did not occur. This causes the model to favor a curve that only slowly increases $c(p)$ as a function of the observed probability estimates.

Table 3 shows that the Recalibrate then Average model had a MPE score of 0.123, which is an 19.7 % improvement over the ULinOP model. The Recalibrate then Average model with Log Odds averaging had a MPE score of 0.112, which is an 26.7 % improvement over the ULinOP model. Therefore, we obtain a greater improvement in model performance by using the log odds averaging approach. For both models, the 95 % confidence interval for the percentage improvement does not contain zero, which indicates that these models are performing significantly better than the basic unweighted average model.

4.3 A hierarchical recalibrate then average model

To evaluate the hierarchical recalibrate then average model, we examined the posterior predictive distribution and the joint posterior distribution of the parameters. Figure 7 shows the posterior predictive distribution for six judges: $i = \{75, 139, 276, 1243, 1393, 1394\}$. To plot these distribution, we took 1,000 samples from the estimated joint posterior distribution for (γ_i, δ_i), and then converted the observed $y_{i,j}$s by using Eq. (2). The figure clearly shows the individual differences between the judges. For example, Judge 1243 has a very different calibration function than Judges 75 and 139. Specifically, when Judge 1243 provided small probability forecasts (e.g., 0.1), the calibration curve pushed this probability upward and when he or she provided large probabilities (e.g., 0.8), the fitted calibration curve pushed this probability downward. However, the opposite pattern occurred for Judges 75 and 139. Figure 7 also shows that some judges were well-calibrated (e.g., Judge 276). We should emphasize that all but one of these judges (i.e., except Judges 1243 and 276) showed clear patterns of subcertain responding and the curves are not symmetrical.

As previously mentioned, one benefit of the hierarchical model is that it includes individual differences, allowing for much more flexible calibrations. Another substantial benefit of this model (and hierarchical Bayesian models in general) is that when there are only a few observations, the model "borrows" power from the estimates of the other judges in the sample. For example, Judges 1393 and 1394 both responded to only one item in the set. As a consequence, the estimates of γ and δ for these judges are reflective of the prior distribution for the group, which is governed by the parameters $\delta_\alpha, \gamma_\alpha, \delta_\beta$, and γ_β. Because these hierarchical parameters are informed by all of the individual estimates for each judge in the

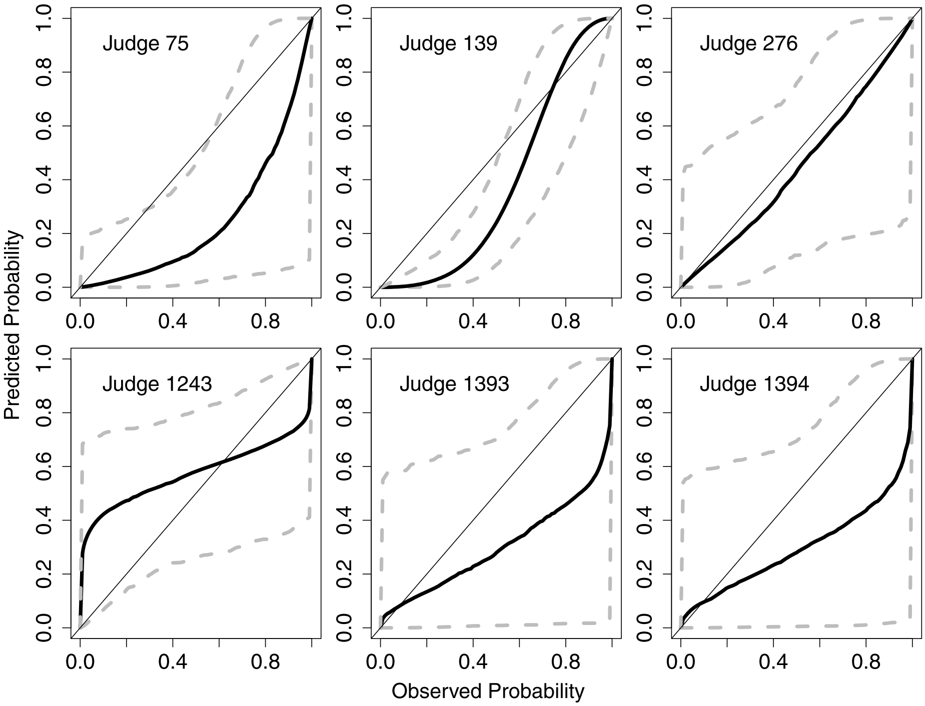

Fig. 7 Illustrative examples for the Hierarchical Recalibrate then Average model. The median (*black lines*) and 95 % credible interval (*gray lines*) for the posterior predictive distribution for nine judges in the data set

group, when there is little information for an individual judge, the model relies more heavily on the prior estimates for these judges.

To further illustrate the range of individual differences across judges, we compared the mean of the posterior distributions for γ_i and δ_i for each judge. The top left and bottom right panels of Fig. 8 show the marginal distributions of γ and δ, respectively, and the top right and bottom left panels of Fig. 8 show the joint distributions of the estimates. Judges 75, 95, 139, and 1243 are shown by the filled square, circle, triangle and diamond, respectively, and Judges 1393 and 1394 are represented by the asterisk symbol. Because Judges 1393 and 1394 have so few observations, their estimates are most similar to the priors, where there is a heavy concentration of estimates (see the marginal distributions of γ and δ in Fig. 8).

The Hierarchical Recalibrate then Average model without log odds averaging obtained a MPE of 0.141, which is a 8 % improvement over the ULinOP. However, the Hierarchical Recalibrate then Average model with Log Odds averaging obtained a MPE of 0.110, which is a 28.2 % improvement over the ULinOP. The model performance of the hierarchical model with log odds averaging is slightly better than the non-hierarchical version. However, this difference is small (and the pair-wise comparisons shown later will not show an advantage for the hierarchical model). Therefore, this suggests that although the modeling of individual differences does not provide a substantial improvement in prediction performance.

4.4 Model comparison at the individual event level

Overall, based on the average prediction error (MPE), it appears that the Hierarchical Recalibrate then Average Log Odds model is the best performing model among all the models

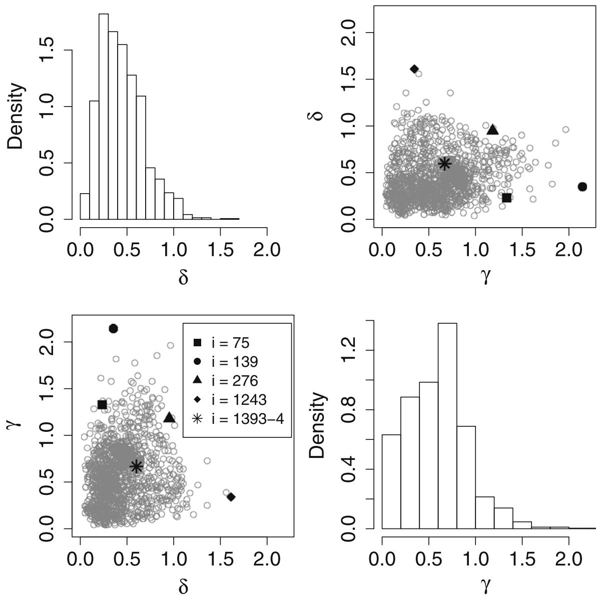

Fig. 8 The marginal (*top left* and *bottom right panels*) and joint (*top right* and *bottom left panels*) distributions of the mean estimates for the 1401 judges in the experiment for the Hierarchical Recalibrate then Average model. Judges 75, 95, 139, and 1243 are shown by the *filled square*, *circle*, *triangle* and *diamond*, respectively, and Judges 1393 and 1394 are represented by the *asterisk symbol*

that we examined. However, this global measure of model prediction error ignores some important patterns at the individual event level (Broomell et al. 2011). For example, one model can achieve a low MPE relative to another model by having substantially better Brier scores on a few forecasting problems, while still doing more poorly on a majority of the problems. The pair-wise model comparison score discussed above emphasizes the latter form of model improvement because it measures the number of forecasting problems on which one model is better than another, regardless of the magnitude of the improvement.

Table 4 shows the pair-wise model comparison scores between six models of interest in the cross validation procedure. Each element in the table shows the number of times the model corresponding to the row fit the data better than the model corresponding to the column, and the percentages are shown in the parentheses. Because the CTALO (Recalibrate then Average Log Odds) model is the only model whose rows contains percentages that are all above 50 %, it is the best performing model in this pair-wise comparison. Importantly, the CTALO model outperforms the hierarchical extension of this model (HCTALO) in 131 out of 176 forecasting problems (74 %).

5 General discussion

In this article, we have examined several models that use the "linear in log odds" function (see Eq. (2)) to recalibrate individual or average judgments and improve the prediction of future events. We found that the order and type of calibration and aggregation had a large impact on the model performance. Overall, in the pair-wise model comparison we found that the Recalibrate then Average Log Odds model has a lower prediction error than any other model on the majority of events. It also performs only slightly worse than the Hierarchical Recalibrate then Average Log Odds model on the average prediction error. Therefore, we conclude that Recalibrate then Average Log Odds model is the simplest and best approach to aggregate forecasting judgments in the presence of systematic biases.

As mentioned in the introduction, in comparing the models we contrasted the consequences of a number of modeling assumptions, including (1) the point at which recalibration occurs (before or after aggregation), (2) the space in which the recalibration and aggregation should take place, (3) the extent to which individual differences are taken into account. Our main goal was to assess how these different assumptions affected the model performance. We now discuss each of these topics in turn and also discuss the influence of the coding scheme for events.

5.1 The order of aggregation and calibration

Our first research question was whether it is better to first aggregate and then recalibrate or recalibrate then average. The answer is that it depends on whether it is probabilities or log-odds that are being averaged. Working solely in the probability scale, we obtained better results in comparisons against the ULinOP when averaging first than when calibrating first (compare rows 2 and 4 in Table 3 and the first two cells under UW in Table 4). However, when averaging log-odds, it is better to calibrate first (row 6 in Table 3 and cell 5 under UW in Table 4).

5.2 The aggregation space

We then compared working with probability and log-odds scales and found that aggregating on the log odds instead of the probability scale led to a reversal in the order of the aggregation and calibration methods to achieve the best results. Specifically, the Recalibrate then Average Log Odds model outperformed the ULinOP by 26.7 %, and achieved the second best MPE score at 0.112. Taking these new results into account, our conclusion above was reversed: it is better to first recalibrate the individual judgments and then aggregate these recalibrated values on the log odds scale.

5.3 The inclusion of individual differences

We examined the benefits of including individual differences in the models. Both of our hierarchical models assigned separate calibration parameters for each individual and estimated the individual parameters in a hierarchical Bayesian approach. This allowed for the estimation of these parameters even for individuals who contributed only a few judgments. For the Hierarchical Recalibrate then Average Log Odds model, we found that taking individual differences into account when aggregating probability judgments led to a 1.5 % improvement in the PDM. However, in the pair-wise comparisons, the hierarchical model performed systematically worse than the non-hierarchical model equivalent (the Recalibrate then Average Log Odds Models). This shows that the any modeling advantage from the hierarchical model comes from improved performance on a small number of forecasting problems and not a systematic improvement across the majority of forecasting problems.

5.4 The role of coding

Finally, we discuss the role of coding. Although the LLO function is not a probability function, we still will work with probabilities in binary forecasts. For example, suppose p is the elicited (and possibly aggregated) probability for the occurrence of some event of interest, we can use $c(p)$ for the transformed probability of an event happening and $1 - c(p)$ for the probability that an event will not happen. Alternatively, we can use a reverse coding scheme where p is the probability for the occurrence of some event, $1 - p$ is the judgment elicited, $1 - c(1 - p)$ for the transformed probability of an event happening, and $c(1 - p)$ for the transformed probability that an event will not happen. The reverse coding scheme will not affect the performance of the models because the transformation still maps to a particular calibration function, specifically

$$\begin{aligned} 1 - c(1 - p|\gamma, \delta) &= \frac{p^\gamma}{\delta(1 - p)^\gamma + p^\gamma} \\ &= c(p|\gamma, 1/\delta). \end{aligned}$$

In the Bayesian framework, for the performance results to be exact, the contribution of the prior for the regular coding scheme must be equivalent to contribution of the prior for the reverse coding scheme for the parameter δ (i.e., the prior for γ can remain the same). For example, for the Average then Calibrate Model we specified that $\delta \sim \Gamma(\alpha, \beta)$, where $\alpha = \beta = 1$. In the reverse coding scheme, an equivalent prior for $1/\delta$ is the inverse gamma distribution, such that $1/\delta \sim \Gamma^{-1}(\alpha, 1/\beta)$. Thus, with the appropriate selection of the prior distribution, the particular coding of the responses and event resolutions has no effect on model performance when using the LLO function.

5.5 Alternatives

There are several other model-based approaches that should be examined and considered in future work. They are important because they provide a model for the internal representation of the judge, which recalibration functions neglect. For example, the Decision Variable Partition model (Ferrell and McGoey 1980) uses signal detection theory as a base representation of the underlying distributions for the alternatives. While this model provides a generally adequate fit to the data (e.g., Suantak et al. 1996), it has been criticized by Keren (1991) for not providing a description of the cognitive processes underlying confidence. However, more recent models have taken this general approach but provide possible explanations of the underlying cognitive processes (e.g., Jang et al. 2012; Wallsten and González-Vallejo 1994).

Other models assume that the judge has a perfectly accurate representation of the observed relative frequency, but due to random error, the probability elicited by the judge is a distorted and usually inaccurate version of the observed relative frequency. Erev et al. (1994) demonstrated through simulation that models of this type can produce typical overconfidence patterns in the data. Other models in this general framework assume the error is attributable to the judgment (the stochastic judgment model; Wallsten and González-Vallejo 1994), aspects of the environment (the ecological error model; Soll 1996) or both (e.g., Juslin and Olsson 1997).

Finally, there exist a variety of mathematical models that explicitly describe psychological processes underlying confidence and/or subjective probability. These include the Poisson race model (Merkle and Van Zandt 2006; Van Zandt 2000), HyGene (Thomas et al. 2008), and the two-stage dynamic signal detection model (Pleskac and Busemeyer 2010).

Although these model-based approaches are important because they offer information about the underlying processes driving the decision, for the sake of simplicity we did not include them in this comparison that focuses primarily on maximizing the predictive accuracy of group forecasts.

6 Conclusions

We have demonstrated that several aggregation methods surpass the predictive accuracy of the ULinOP, at least in the limited situation of forecasting whether or not the status quo will change within a specified time frame. Our best performing model first corrected for systematic distortions and then aggregated the calibrated judgments on the log odds scale. The hierarchical version of the calibration model, which allowed for individual differences in the nature of the systematic distortion, outperformed its single-level counterpart, but this difference was small; and we concluded for this study does not merit the additional complexity involved.

The present work builds on a growing body of literature evaluating various calibration and aggregation methods. Aggregating multiple subjective probability estimates to improve the performance of the estimate ties into the concept of the "wisdom of the crowd" effect (Surowiecki 2004), which has usually been studied in the context of a single magnitude estimate. For example, Galton (1907) found that when people were asked to estimate the weight of a butchered ox, the average of these estimates was almost exactly correct, despite wide variability in the estimates. More recent studies have demonstrated this effect in more complicated situations such as optimizing solutions in combinatorial problems (Yi et al. 2010; Yi et al. 2011), inferring expertise (Lee et al. 2011), maximizing event recall accuracy (Hemmer et al. 2011) and solving ordering problems (Miller et al. 2009; Steyvers et al. 2009). The present work suggests substantially greater improvement can be attained by taking systematic distortions in the individual judgments into account. The best aggregation performance might be obtained with models that first recalibrate individual estimates and then combine these recalibrated judgments on an appropriate scale.

Acknowledgements Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20059. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

Appendix

In each of the JAGS codes below, we pass the program a series of data structures to facilitate the estimations. For all of the single-level models, we let y contain all probability judgments, so it is a vector containing all of the judgments for the first item, followed by the second item, and so on. To tell the program when the responses change from one item to the next, we pass it a vector containing these indexes, called $nresps$. For example, $nresps[1]$ might equal 740, which indicates that all of the responses from $y[1]$ to $y[740]$ are for the first item. For the Average then Recalibrate model, y contains only the averages of all judgments for each item, so $nresps$ is not needed. For the hierarchical models, y is a matrix where the rows correspond to the observation, the first column contains the subject index eliciting the judgment, and the second column contains the judgment. In addition, we pass JAGS a new variable $xstar$, which contains the event resolution information for each individual separately (see text above).

Finally, we pass the program the upper indexes for each loop. We let nd be the total number of probability judgments, nq be the number of items in the set, nk is the number of items with known outcomes, and ns is the number of subjects (only used for the hierarchical models). For the hierarchical models, nk is the number of responses to items having known outcomes (i.e., nk is the length of the vector $xstar$).

A.1 Calibrate then average

```
##### Recalibrate Then Average
model{
# Specify the recalibration function
for(k in 1:nd){
p[k] <-delta*(y[k])^gamma/(delta*(y[k])^gamma+(1-y[k])^gamma)
}
# Collapse the information in the p matrix through aggregation
mu[1] <- mean(p[1:nresps[1]])
for(j in 2:nq){
mu[j] <- mean(p[(1+nresps[j-1]):nresps[j]])
}
# The resolution is a result of the latent mean parameters
for(j in 1:nk){
x[j] ~ dbern(mu[j])
}
# Priors on model parameters
delta ~ dgamma(1,1)
gamma ~ dgamma(1,1)
}
```

A.2 Calibrate then average log odds scale

```
##### Recalibrate Then Average Log Odds Scale
model{
# Specify the recalibration function
for(k in 1:nd){
p[k] <- gamma*log(y[k]/(1-y[k])) + log(delta)
}
# Collapse the information in the p matrix through aggregation
mu_star[1] <- mean(p[1:nresps[1]])
mu[1] <- exp(mu_star[1])/(1+exp(mu_star[1]))
for(j in 2:nq){
mu_star[j] <- mean(p[(1+nresps[j-1]):nresps[j]])
mu[j] <- exp(mu_star[j])/(1+exp(mu_star[j]))
}
# The resolution is a result of the latent mean parameters
for(j in 1:nk){
x[j] ~ dbern(mu[j])
}
# Priors on model parameters
delta ~ dgamma(1,1)
gamma ~ dgamma(1,1)
}
```

A.3 Average then recalibrate

```
##### Average then Recalibrate
model{
# Specify recalibration function
for(k in 1:nq){
p[k] <- delta*(y[k])^gamma/(delta*((y[k])^gamma)+(1-y[k])^gamma)
}
# The resolution is a result of the latent mean parameters
for(j in 1:nk){
x[j] ~ dbern(p[j])
}
# Priors on model parameters
delta ~ dgamma(1,1)
gamma ~ dgamma(1,1)
}
```

A.4 Hierarchical recalibrate then average

```
##### Hierarchical Recalibrate then Average
model{
# Specify recalibration function
for(k in 1:nd){
p[k] <- delta[y[k,1]]*(y[k,2])^gamma[y[k,1]]/
(delta[y[k,1]]*(y[k,2])^gamma[y[k,1]]+(1-y[k,2])^gamma[y[k,1]])
}
# The resolution is a result of the latent mean parameters
for(j in 1:nk){
xstar[j] ~ dbern(p[j])
}
# Priors on individual-level model parameters
for(i in 1:ns){
delta[i] ~ dgamma(delta_mu,delta_sigma)
gamma[i] ~ dgamma(gamma_mu,gamma_sigma)
}
# Priors on hyperparameters
delta_mu ~ dgamma(1000,1000)
gamma_mu ~ dgamma(1000,1000)
delta_sigma ~ dgamma(1000,1000)
gamma_sigma ~ dgamma(1000,1000)
}
```

A.5 Hierarchical recalibrate then average log odds

```
##### Hierarchical Recalibrate then Average Log Odds
model{
# Specify recalibration function
for(k in 1:nd){
p[k] <- gamma[y[k,1]]*log(y[k,2]/(1-y[k,2])) + log(delta[y[k,1]])
}
# Aggregate the p vector
mu_star[1] <- mean(p[1:nresps[1]])
mu[1] <- exp(mu_star[1])/(1+exp(mu_star[1]))
for(j in 2:nq){
mu_star[j] <- mean(p[(1+nresps[j-1]):nresps[j]])
mu[j] <- exp(mu_star[j])/(1+exp(mu_star[j]))
}
# The resolution is a result of the latent mean parameters
for(j in 1:nk){
xstar[j] ~ dbern(mu[j])
}
# Priors on individual-level model parameters
for(i in 1:ns){
delta[i] ~ dgamma(delta_mu,delta_sigma)
gamma[i] ~ dgamma(gamma_mu,gamma_sigma)
}
# Priors on hyperparameters
delta_mu ~ dgamma(1000,1000)
gamma_mu ~ dgamma(1000,1000)
delta_sigma ~ dgamma(1000,1000)
gamma_sigma ~ dgamma(1000,1000)
}
```

References

Ariely, D., Au, W. T., 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*, 130–147.

Arkes, H. R., Dawson, N. V., Speroff, T., Harrell, F. E. Jr., Alzola, C., Phillips, R., Desbiens, N., Oye, R. K., Knaus, W., & Connors, A. F. Jr. (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making*, *15*, 120–131.

Armstrong, J. S. (2001). *Principles of forecasting*. Norwell: Kluwer Academic.

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes*, *67*, 91–110.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: a critical examination. *Organizational Behavioral and Human Decision Processes*, *65*, 212–219.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Broomell, S. B., Budescu, D. V., & Por, H. H. (2011). Pair-wise comparison of multiple models. *Judgement and Decision Making*, *6*, 821–831.

Brown, S., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Budescu, D. V., & Johnson, T. R. (2011). A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making*, *6*, 857–869.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, *2*, 159–225.

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

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. Boca Raton: CRC Press.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology. Human Perception and Performance*, *7*, 928–935.

Clemen, R. T. (1986). Calibration and the aggregation of probabilities. *Management Science*, *32*, 312–314.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography (with discussion). *International Journal of Forecasting*, *5*, 559–583.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*, 39–46.

Cooke, R. M. (1991). *Experts in uncertainty: opinion and subjective probability in science*. New York: Oxford University Press.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes*, *26*, 32–53.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Machine learning: Proceedings of the thirteenth international conference* (pp. 325–332). San Francisco: Morgan Kauffman.

Fryback, D. G., & Erdman, H. (1979). Prospects for calibrating physicians' probabilistic judgments: design of a feedback system. In *Proceedings of the IEEE international conference on cybernetics and society*.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. New York: Chapman and Hall.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*, 129–166.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Härdle, W. (1992). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Hemmer, P., Steyvers, M., & Miller, B. J. (2011). The wisdom of crowds with informative priors. In *Proceedings of the 32rd annual conference of the cognitive science society*. Lawrence Erlbaum.

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, *50*, 597–604.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*, 186–200.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: a sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344–366.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Karmarker, U. S. (1978). Subjectively weighted utility: a descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, *21*, 61–72.

Keren, G. (1991). *Calibration and Probability Judgments: Conceptual and Methodological Issues*. *Acta Psychologica*.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lee, M. D., Steyvers, M., de Young, M., & Miller, B. J. (2011). A model-based approach to measuring expertise in ranking tasks. In *Proceedings of the 33rd annual conference of the cognitive science society*. Cognitive Science Society.

Lee, M. D., & Wagenmakers, E.-J. (2012). A course in Bayesian graphical modeling for cognitive science. Available from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf; last downloaded January 1, 2012.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

Merkle, E. C. (2010). Calibrating subjective probabilities using hierarchical Bayesian models. In S.-K. Chai, J. J. Salerno, & P. L. Mabry (Eds.), *Lecture notes in computer science: Vol. 6007*. *Social computing, behavioral modeling, and prediction (SBP) 2010* (pp. 13–22).

Merkle, E. C., Smithson, M., & Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, *55*, 57–67.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology. General*, *135*, 391–408.

Miller, B. J., Hemmer, P., Steyvers, M., & Lee, M. D. (2009). The wisdom of crowds in ordering problems. In *Proceedings of the ninth international conference on cognitive modeling*, Manchester, UK.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Murphy, A. H., & Winkler, R. L. (1974). Credible interval temperature forecasting: some experimental results. *Monthly Weather Review*, *102*, 784–794.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*, 489–500.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. Hoboken: Wiley.

Page, L., & Clemen, R. T. (2012). Do prediction markets produce well calibrated probability forecasts? Manuscript in preparation.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection theory: a theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: the roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*, 117–137.

Staël von Holstein, C. A. S. (1970). Measurement of subjective probability. *Acta Psychologica*, *34*, 146–159.

Steyvers, M., Lee, M. D., & Miller, B. J. (2009). Wisdom of crowds in the recollection of order information. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems.* (Vol. 22, pp. 1785–1793). New York: MIT Press.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201–221.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*, 269–283.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representations of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 582–600.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Science*, *29*, 151–173.

Wallsten, T. S., Budescu, D. V., & Erev, I. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: a stochastic model of judgment and response. *Psychological Review*, *101*, 490–504.

Warnaar, D., Merkle, E., Steyvers, S., Wallsten, T., Stone, E., Budescu, D., Yates, J., Sieck, W., Arkes, H., Argenta, C., Shin, Y., & Carter, J. (2012). The aggregative contingent estimation system: selecting, rewarding, and training experts in a wisdom of crowds approach to forecasting. In *2012 AAAI symposium*. on Wisdom of the Crowd.

Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, *52*, 165–174.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology*, *23*, 219–224.

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

Yates, J. F. (1982). External correspondence: decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132–156.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs: Prentice Hall.

Yi, S. K. M., Steyvers, M., & Lee, M. D. (2011). *The wisdom of crowds in combinatorial problems*. Cognitive Science.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. (2010). Wisdom of crowds in minimum spanning tree problems. In *Proceedings of the 32nd annual conference of the cognitive science society*. Lawrence Erlbaum.

Zhang, H., & Maloney, L. T. (2011). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14.