Subjective Probability Correction for Uncertainty Representations

Fumeng Yang
fy@northwestern.edu
Northwestern University
Evanston, IL, USA

Maryam Hedayati
maryam.hedayati@u.northwestern.edu
Northwestern University
Evanston, IL, USA

Matthew Kay
mjskay@northwestern.edu
Northwestern University
Evanston, IL, USA

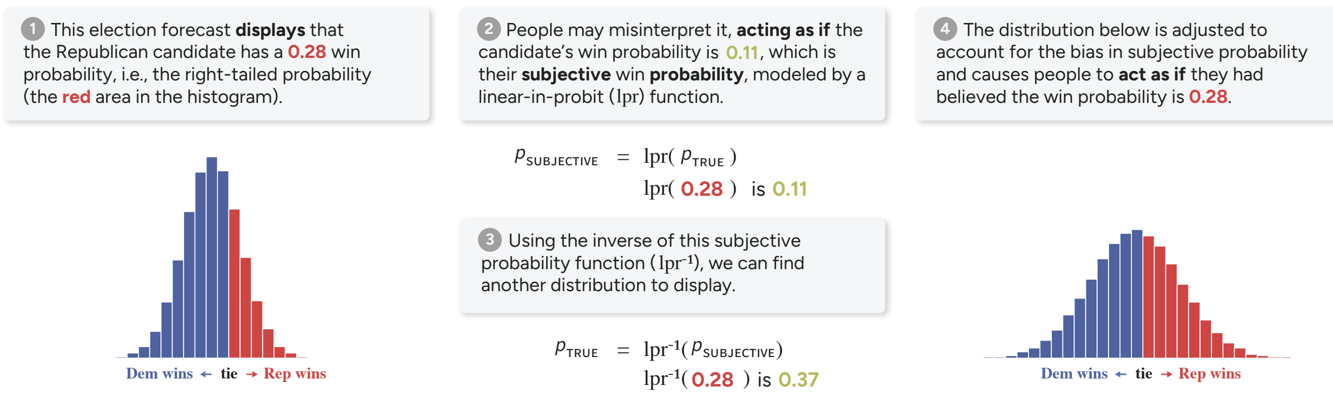


Figure 1: The concept of subjective probability correction: In this exemplar election forecast, the right-tailed probability represents the Republican candidate's win probability. (1) When viewing a win probability of 0.28, (2) people may misinterpret it and *act as if* the candidate has a 0.11 probability of winning. (3) To compensate for this bias in decision-making, we can use the inverse of the subjective probability function, which allows us to start with the desired probability, say 0.28, and find another distribution to display. (4) The resulting bias-corrected distribution causes people to *act as if* their subjective probability of that candidate winning is the desired 0.28, while actually displaying a win probability of 0.37.

ABSTRACT

We propose a new approach to uncertainty communication: we keep the uncertainty representation fixed, but adjust the distribution displayed to compensate for biases in people's subjective probability in decision-making. To do so, we adopt a linear-in-probit model of subjective probability and derive two corrections to a Normal distribution based on the model's intercept and slope: one correcting all right-tailed probabilities, and the other preserving the mode and one focal probability. We then conduct two experiments on U.S. demographically-representative samples. We show participants hypothetical U.S. Senate election forecasts as text or a histogram and elicit their subjective probabilities using a betting task. The first experiment estimates the linear-in-probit intercepts and slopes, and confirms the biases in participants' subjective probabilities. The second, preregistered follow-up shows participants the bias-corrected forecast distributions. We find the corrections substantially improve participants' decision quality by reducing the integrated absolute error of their subjective probabilities compared to the true probabilities. These corrections can be generalized to any univariate probability or confidence distribution, giving them broad applicability. Our preprint, code, data, and preregistration are available at `https://doi.org/10.17605/osf.io/kcwxm`

preprint

CCS CONCEPTS

• **Human-centered computing** → *Visualization design and evaluation methods*; *Information visualization*; *Empirical studies in visualization*; **User models**.

KEYWORDS

uncertainty visualization, subjective probability, perception, election forecasts

ACM Reference Format:
Fumeng Yang, Maryam Hedayati, and Matthew Kay. 2023. Subjective Probability Correction for Uncertainty Representations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3580998

1 INTRODUCTION

Subjective probability measures the quality of decisions made under uncertainty [2, 36]. It is the internal probabilities people *act as if* they had believed when making decisions. In uncertainty communication, one way to improve subjective probability is to assume a probability distribution over future events, then tackle how to represent this distribution in a way that reduces biases in subjective probabilities, bringing them closer to the true probabilities being communicated. This has been a fruitful line of inquiry in uncertainty visualization, leading to visualization types that improve decision quality [12, 27, 41]. However, there may exist a limit to how much we can improve decision quality by modifying representations alone; for example, some improved representations may only increase decision quality for people with higher working memory capacity [42].

We introduce a new approach that fixes the uncertainty representation, but **adjusts the distribution being displayed** to account for biases in subjective probability. Intuitively, we must "undo" the distortions that occur when transitioning from true probability to subjective probability. More formally, if showing people distribution X will cause them to act as if they had seen some other distribution, say $g(X)$, then we need an invertible function g that describes people's subjective probabilities as a function of the true probabilities. We then invert g and display the distribution $X' = g^{-1}(X)$, so people will act as if they had seen X. This adjustment to the displayed probability distribution **compensates for biases** in subjective probability to improve decision quality, and we call it a **subjective probability correction**.

To create a subjective probability correction, we adopt a linear-in-probit model of subjective probabilities, a mathematically convenient variation on the linear-in-log-odds model [61] that generalizes both prospect theory [26] and models of proportion perception [15, 22]. In principle, the linear-in-probit model can be used to adjust any univariate distribution. We demonstrate how, for Normal distributions, we can scale and shift that distribution based on the intercept and slope of the linear-in-probit model to obtain a bias-corrected distribution. As this correction may move the mode of the distribution, we also present another correction that uses the skew-Normal distribution to preserve the mode of the distribution and one focal probability.

We evaluate our proposed corrections in the context of U.S. Senate election forecasts. These forecasts predict candidates' (or parties') vote percentages and compute win probabilities from the vote percentage distributions. In recent years, U.S. election forecasts have become controversial partly because people tend to misinterpret these probabilities [11, 58], making them a promising testbed.

Specifically,

(1) We derive two corrections for Normal distributions based on a linear-in-probit model of subjective probability: a **Normal correction** and a **skew-Normal correction**. These corrections can be applied so long as the intercept and slope of the linear-in-probit model for a given decision task are known.

(2) We conduct an online experiment using a Senate election scenario and a U.S. demographically-representative sample (N=306), and test two common representations (**text** and **histogram**) for the forecast distributions. We elicit participants' subjective probabilities of a candidate winning under a betting task, estimate the linear-in-probit intercepts and slopes for this task, and measure integrated absolute error of subjective probabilities compared to the true probabilities.

(3) We derive bias-corrected forecast distributions from the estimated intercepts and slopes, and, in a preregistered follow-up, we repeat the experiment (N=603) but show participants text and histograms of these bias-corrected distributions. The corrections substantially **improve decision quality**. For example, the skew-Normal correction **reduces 60% of integrated absolute error for text** (the posterior median reduces from 0.13 [0.12, 0.14] to 0.054 [0.030, 0.076]) and **30% for histograms** (the posterior median reduces from 0.092 [0.074, 0.11] to 0.064 [0.045, 0.081]), bringing subjective probabilities much closer to the true probabilities. The corrections debias the linear-in-probit intercepts, but do not completely debias the linear-in-probit slopes.

While our approach substantially improves decision quality, our inability to fully correct biases in subjective probability opens up avenues for future work. Perhaps there is a ceiling on how much we can improve, or perhaps considering a mixture of decision strategies people use would allow a more complete correction [27]. In any case, an error reduction of 5–10 percentage points is on par with improvements seen by modifying uncertainty representations [27, 31], suggesting our subjective probability corrections may be a valuable tool in the toolbox for uncertainty communication, complementing work on improved uncertainty representations.

Preregistration statement Our first experiment does not have a preregistration, because we do not have any expectation of effect size nor any specific hypotheses. We use the data from this experiment to decide on model specification, priors, and sample size for the second experiment; these are preregistered. We also preregistered three measures for the second experiment: the (1) intercept and (2) slope of the linear-in-probit model, and (3) integrated absolute error of subjective probability elicited from participants' decision-making.

2 BACKGROUND

Our works draw upon three areas: (1) subjective probability in decision-making, (2) uncertainty visualization, and (3) perceptual optimization for visualization.

2.1 Subjective probability in decision-making under uncertainty

Subjective probability is commonly used in decision analysis [19]. This concept is different but related to the true (objective) probability used by statisticians. It describes a decision-maker's underlying belief in a probabilistic outcome, which they use for estimating their utilities or expected rewards [19, 36]. It is a behavioral measure, often inferred from the choices people make in a sequence of lotteries with an incentive [45, 56].

By contrast, directly reported probabilities of visually perceived proportions, which are sometimes used to evaluate uncertainty visualizations, do not measure decision quality [23]. For example, people may accurately repeat back the exact probability of rain from a weather forecast presented as text or a histogram. However, their responses might not match their underlying belief in the probability of rain (their subjective probability), perhaps better measured by whether or not they actually bring an umbrella. In an election forecasting context, the subjective probability of a candidate winning may drive voters to cast a ballot or mobilize in their community [16] or lead them to be surprised when a candidate with a 0.3 forecasted win probability ultimately wins an election.

In these decision-making processes, people do not usually perform mental calculations, but instead rely on cues or heuristics [7], leading to some distortion of judgment or misperception of probabilities, which are *biases* in their subjective probability [7]. Our work builds upon such literature to measure people's subjective probability in decision-making and aims to correct for people's biases to improve their decision quality.

2.2 Uncertainty visualization

Much work in uncertainty visualization attempts to find more effective representations of distributional uncertainty, e.g., through encodings based on intervals [8, 12, 31], density functions [8, 12, 21, 25, 31], or cumulative distribution functions [12, 25, 59]; or by employing frequency-framing approaches such as quantile dotplots [12, 27, 31], hypothetical outcome plots (HOPs) [24, 28], or spaghetti plots [33, 46]. Often this work is grounded in attempts to help the viewer better understand uncertainty or make better decisions from a representation. For example, Helske et al. [21] used densities with faded tails in an attempt to reduce researchers' reliance on dichotomous thinking; frequency representations (like dotplots [31] or HOPs [24]) are also commonly used to improve decisions under uncertainty, inspired by research in cognitive science that suggests people reason better with discrete outcomes than continuous probabilities [17]. Much of this work fundamentally rests on visualizing distributions or summary statistics of distributions (whether they are probability distributions or confidence distributions [60]), and so is compatible with our approach to subjective probability correction through adjusting distributions. A related bias-correction approach for uncertainty visualization is Correll et al.'s Value-Suppressing Uncertainty Palettes [10]: they merge successive categories in a bivariate colormap and suppress the color of the point estimate when uncertainty is larger. Their approach is more heuristic and is not based on models of perception or decision-making. Our model-driven approach could provide a theoretical grounding for similar value-suppression functions, as well as a principled way to choose *how much* value to suppress [30].

2.3 Perceptual optimization for visualization

Perceptual optimization and bias correction are commonly employed in other areas of visualization, with a focus on perceptual features. For example, Micallef et al. optimized parameters like opacity for scatterplots based on task objectives [37]. Other examples include adjusting orientation to compensate for biases in trend estimation in scatterplots [34] and including annotations when two bars are perceptually indistinguishable [35]. Color is also oft-targeted for debiasing; for example, previous work constructed perceptually rather than mathematically uniform colormaps (e.g., *viridis* [52] and its corrected version *cviridis* [39]); colormaps have also been optimized for separating different classes in scatterplots [57], differentiating common mark types [55], or highlighting unexpected events [9]. Area is another visual channel amenable to debiasing: Flannery [13] proposed scaling points on maps according to a power-law transformation of area perception rather than raw areas (cf. Stevens' power law [54]). Other approaches changed sampling methods [44] to create perceptually-optimized scatterplots or used a neural network simulation of early perceptual processing to adjust the parameterization of flow visualizations [43]. Our work attempts to bring this tradition to uncertainty visualization by systematically adjusting a displayed distribution using models of people's subjective probability [61].

3 BIAS-CORRECTING PROBABILITY DISTRIBUTIONS

This section describes our mathematical derivation of the subjective probability correction. To help readers follow our derivation, we first introduce the statistical concepts needed using an example of election forecasts.

3.1 Preliminaries: Probability distributions and election forecasts

Forecasts are often made using probability distributions over possible outcomes. For example, election forecasters may show a probability distribution for the vote percentage that a candidate is predicted to receive using 1 a **probability density function** (PDF). The PDF for distribution X is denoted $f_X(x)$. The highest point on the PDF, 2 the **mode**, is the most likely vote percentage the candidate will receive. On a PDF, probability is read by looking at the area under the curve.

In election forecasts, a meaningful 3 **focal point** is 50% vote share: above this value, the candidate wins. In this example, the area under the curve to the right of 50% is 0.3 of the total area under the curve. In other words, 4 the **right-tailed probability**, $\mathbb{P}(X > 50\%)$, is 0.3, which is the candidate's win probability.

The 5 **complementary cumulative distribution function (CCDF)** gives all of the right-tailed probabilities for a distribution.[1] It is denoted $1 - F_X(x) = \mathbb{P}(X > x)$. For example, if we read

[1]Similarly, the cumulative distribution function (CDF), $F_X(x) = \mathbb{P}(X \leq x)$, gives left-tailed probabilities.

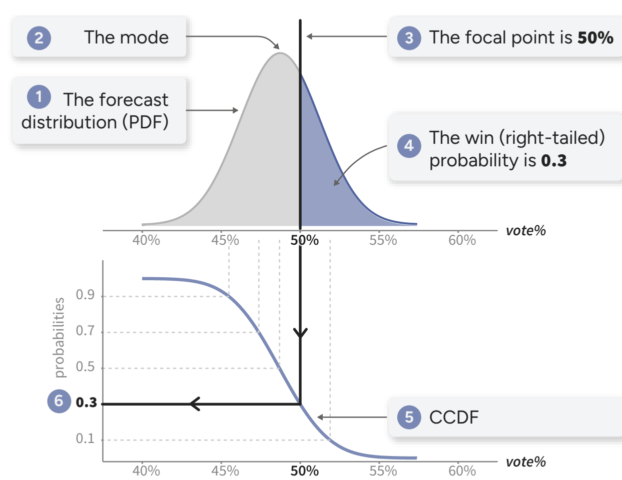


Figure 2: An election forecast (top) and its CCDF (bottom).

down from 50% to the CCDF and across, we see that the probability of winning, $\mathbb{P}(X > 50\%)$, is 6 **0.3**. Our goal then is to translate all of the probabilities in a forecast distribution (e.g., its CCDF) into a new, bias-corrected distribution—one which causes people's subjective probabilities to resemble the original distribution. For that translation, we turn to models of probability perception.

3.2 Linear-in-probit model for subjective probability

In a review of work on subjective probability and proportion perception (including prospect theory [26] and visual perception [15, 22]), Zhang and Maloney [61] propose a **linear-in-log-odds (llo) model** as a good fit for patterns of probability and frequency distortions in a variety of domains. This makes it a promising foundation for a robust bias correction for subjective probability. While their model is expressed in terms of a slope and a crossover point, we express it as a slope (β') and intercept (α'):

$$p_{\text{SUBJECTIVE}} = \text{llo}(p_{\text{TRUE}}) = \text{logit}^{-1}\left[\alpha' + \beta' \cdot \text{logit}(p_{\text{TRUE}})\right] \quad (1)$$

For mathematical convenience, we will use a **linear-in-probit (lpr) model** instead:

$$p_{\text{SUBJECTIVE}} = \text{lpr}(p_{\text{TRUE}}) = \text{probit}^{-1}\left[\alpha + \beta \cdot \text{probit}(p_{\text{TRUE}})\right] \quad (2)$$

The logit and probit functions are both S-shaped functions, and are difficult to distinguish empirically (see Appx. B); consequently, one or the other is often adopted for mathematical convenience [1]. Here, the probit formulation is useful because the probit function is the inverse cumulative distribution function of the standard Normal distribution ($\text{probit}(p) = \Phi^{-1}(p)$), which will allow us to derive a closed-form bias correction for Normal distributions.

The linear-in-probit function is controlled by its intercept (α) and slope (β), which determine the shape of the relationship between true and subjective probabilities (Fig. 3). When both probabilities are probit-transformed, their relationship is linear (Fig. 3b). This model allows for an overall bias (determined by α) in subjective probability, and for distortions which pull probabilities towards 0 or 1 ($\beta > 1$) or towards the center ($\beta < 1$; Fig. 3, the third column).

Besides empirical work suggesting its broad applicability [61], this model has face validity when applied to an election forecasting scenario. Journalists encounter challenges communicating uncertainty in their forecasts to ensure readers do not ignore it [11]. People tend to "round" forecasted win probabilities towards 0 or 1, e.g., misinterpreting a forecast that a candidate has a 0.3 chance of winning as a very unlikely event (Fig. 1), then being frustrated if the candidate ultimately wins the election. This phenomenon can be captured by a linear-in-probit model with $\beta > 1$: large probabilities are pulled towards 1, and small probabilities are pulled towards 0 (Fig. 3, the fourth column).

3.3 Generic bias correction

If we wish for people's subjective probability distribution to be X, we need to display an alternative distribution—the bias-corrected version—such that people will act as if they had seen X. This requires knowing the intercept (α) and slope (β) of the linear-in-probit model, which are domain-dependent [61] and must be empirically measured (Sec. 4). We must also know the probabilities of interest to the viewer: they might be interested in left- or right-tailed probabilities, e.g., loss or win probabilities of a candidate. Assuming we know α and β, and the viewer is interested in left-tailed probabilities ($\mathbb{P}(X \leq x)$), one approach would be to use the inverse of the linear-in-probit function[2] (lpr^{-1}) to transform the cumulative distribution function (CDF) of X to derive a bias-corrected distribution. We call this distribution $X^{<}$, a **left-tailed bias correction**:

$$\underbrace{\mathbb{P}(X^{<} \leq x)}_{\textit{probability we display}} = \text{lpr}^{-1}\big(\underbrace{\mathbb{P}(X \leq x)}_{\textit{people's subjective probability}}\big)$$

Alternatively, the viewer may be interested in right-tailed probabilities ($\mathbb{P}(X > x)$). In election forecasts, this could be the probability a candidate gets more than 50% of the vote and wins the election. Thus, we may use the complementary cumulative distribution function (CCDF) of X to derive $X^{>}$, the **right-tailed bias correction**:

$$\underbrace{\mathbb{P}(X^{>} > x)}_{\textit{probability we display}} = \text{lpr}^{-1}\big(\underbrace{\mathbb{P}(X > x)}_{\textit{people's subjective probability}}\big)$$

Applied to a general distribution, such a correction may require the use of numerical differentiation to find corresponding densities. However, applied to a Normal distribution, we can derive the correction analytically.

3.4 Normal correction for subjective probabilities

If $X \sim \text{Normal}(\mu, \sigma^2)$, then the complementary cumulative distribution function (CCDF) of X gives the probability the candidate receives more than any given vote percentage, and it is denoted:

$$\mathbb{P}(X > x) = 1 - F_{\text{Normal}}\left(x \,\middle|\, \mu, \sigma^2\right) = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{\mu - x}{\sigma}\right)$$

Then given the intercept (α) and slope (β) of the linear-in-probit function, we can derive the **right-tailed bias-corrected** distribution, $X^{>}$, by substituting the CCDF into the inverse of the linear-in-probit function:

$$\begin{aligned}
\underbrace{\mathbb{P}(X^{>} > x)}_{\textit{probability we display}} &= \text{lpr}^{-1}\big(\underbrace{\mathbb{P}(X > x)}_{\textit{people's subjective probability}}\big) \\
&= \text{lpr}^{-1}\left(\Phi\left(\frac{\mu - x}{\sigma}\right)\right) \\
&= \Phi\left(\frac{\mu - \alpha\sigma - x}{\beta\sigma}\right) \\
&= 1 - F_{\text{Normal}}\left(x \,\middle|\, \mu - \alpha\sigma, (\beta\sigma)^2\right) \\
\implies X^{>} &\sim \text{Normal}\left(\mu^{>}, \overset{>}{\sigma}{}^2\right) \\
\text{where } \mu^{>} &= \mu - \alpha\sigma \\
\text{and } \sigma^{>} &= \beta\sigma
\end{aligned}$$

We can similarly derive a **left-tailed bias-corrected** distribution,[3] $X^{<}$:

[2] By inverting Eq. 2, we get $\text{lpr}^{-1} = \text{probit}^{-1}\left(\frac{\text{probit}(x) - \alpha}{\beta}\right) = \Phi\left(\frac{\Phi^{-1}(x) - \alpha}{\beta}\right)$.

[3] The step-by-step derivation is provided in Appx. C.

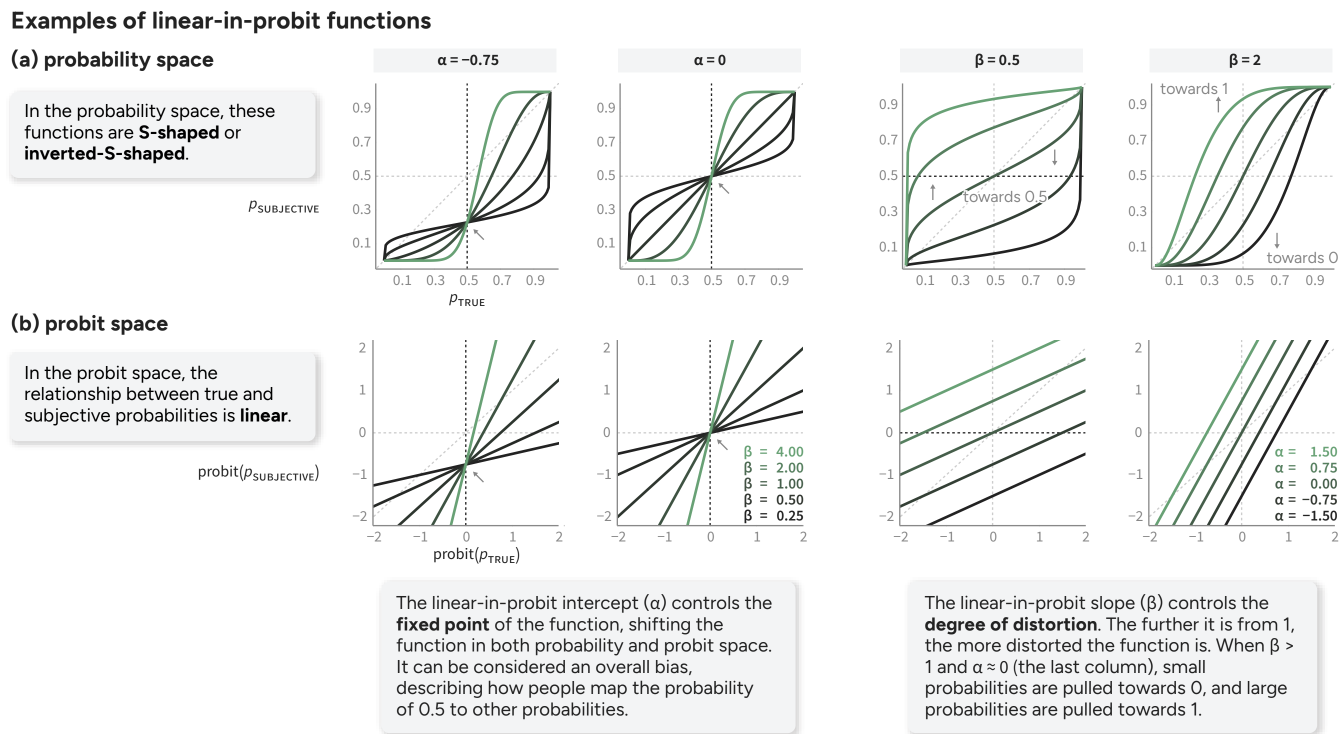


Figure 3: Examples of linear-in-probit models of subjective probability as a function of the true probability. Each panel shows a fixed intercept (α; an overall bias) or a fixed slope (β; the degree of distortion). The bottom row shows the same models as in the top row, with both the $x-$ and $y-$axis transformed by the probit function. More examples are provided in Appx. A.

probability we display *people's subjective probability*

$$\begin{aligned} \mathbb{P}(X^< \leq x) &= \text{lpr}^{-1}\left(\mathbb{P}(X \leq x)\right) \\ \implies X^< &\sim \text{Normal}\left(\mu^<, {\sigma^<}^2\right) \\ \text{where } \mu^< &= \mu + \alpha\sigma \\ \text{and } \sigma^< &= \beta\sigma \end{aligned}$$

This suggests that when faced with Normally-distributed uncertainty, a viable correction for subjective probability is to scale the original distribution by β and shift it up by $\alpha\sigma$ (left-tailed correction) or down by $\alpha\sigma$ (right-tailed correction; see Fig. 5a).

One limitation of this correction is that when α is nonzero, the mode of the distribution is also shifted. This may not be desirable: in the context of election forecasting, for example, if the forecast is for the vote percentage in a two-party race, the modal prediction may be shifted from below 50% of the vote to above it, changing which party is forecast to win in the most likely case (see Fig. 5). It may be desirable to keep the predicted winner unchanged, which motivates another correction method as below.

3.5 Skew-Normal correction to preserve modal forecast

If we wish to preserve the modal probability when the point prediction is meaningful, we cannot use the Normal correction, as it will shift the mode of the distribution. Since the Normal correction is entailed by a transformation of all right- (or left-) tailed probabilities, we must relax those conditions. We will derive a **right-tailed skew-Normal correction**, X^*, with the following conditions:[4]

(1) Instead of ensuring all right-tailed true probabilities are accurately reflected by their corresponding subjective probabilities, we will **preserve a focal probability**; i.e., we want $\mathbb{P}(X^* > x_{\text{FOCAL}}) = \text{lpr}^{-1}(\mathbb{P}(X > x_{\text{FOCAL}}))$ for some domain-specific x_{FOCAL}. In the election forecasting scenerio, we preserve $\mathbb{P}(X > 50\%)$; i.e., the probability that one candidate gets more than 50% of the vote and wins the election.
(2) Unlike with the Normal correction, we will **preserve the mode of the distribution**; i.e., we want $\text{mode}(X^*) = \text{mode}(X)$. Since the mode and mean of a Normal distribution are equal, this implies we want $\text{mode}(X^*) = \mu$.
(3) Finally, so that the width of X^* roughly approximates the corrected Normal distribution apart from the skew, we will **preserve the standard deviation**; i.e., we let $\sigma^* = \sigma^>$.

Unfortunately, there is no closed-form parameterization of the skew-Normal distribution in terms of its mode,[5] so we use numerical optimization to find a skew-Normal distribution with the desired

[4] A left-tailed skew-Normal correction could be derived analogously.
[5] Though it is unimodal, and given a skew-Normal distribution, it is straightforward to use numerical optimization to find the mode by finding the x value that maximizes its density function.

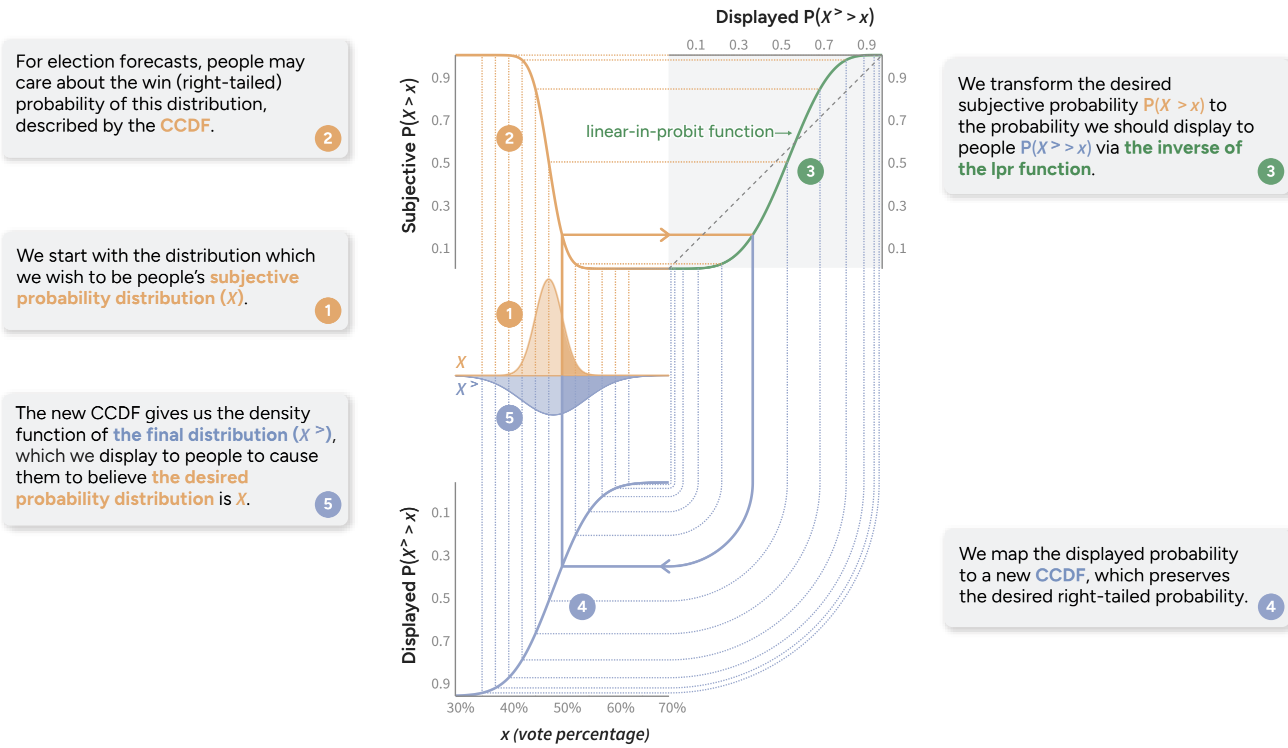


Figure 4: Illustration of the Normal correction for right-tailed probabilities. If we want the 1 orange distribution X to be people's subjective probability distribution, we display the 5 blue distribution $X^{>}$, and subjective quantiles from this distribution are preserved (have the same x values) when translating the two CCDFs (2 4) through the 3 linear-in-probit function.

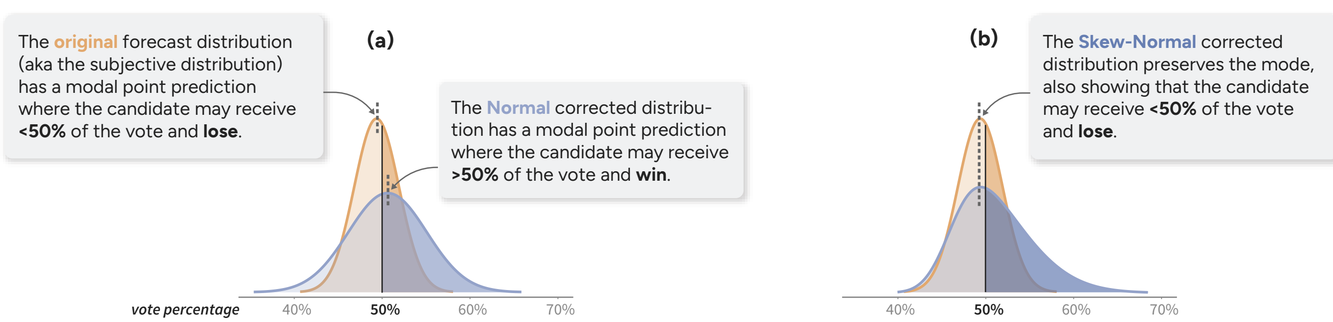


Figure 5: The difference between the Normal and Skew-Normal corrections. (a) The Normal correction may shift the mode of the distribution, changing the modal point prediction; (b) the Skew-Normal preserves the mode of the distribution and the modal point prediction.

mode, focal probability ($\mathbb{P}(X^* > x_{\text{FOCAL}}) = p_{\text{FOCAL}}$), and standard deviation (σ^*).

The typical parameterization of the skew-Normal is defined by location (ξ), scale (ω), and skew (λ): [6]

$$X^* \sim \text{skew-Normal}(\xi, \omega, \lambda) \tag{3}$$

We reparameterize the skew-Normal distribution in terms of its standard deviation (σ^*) to satisfy the third condition:

[6] We use the typical definition of a skew-Normal distribution, with density function:

$$f_{\text{skew-Normal}}(x|\xi,\omega,\lambda) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\left(\frac{x-\xi}{\omega}\right)\right)$$

$$X^* \sim \text{skew-Normal}'(\xi, \sigma^*, \lambda) \quad (4)$$

$$\text{where } \omega = \frac{\sigma^*}{\sqrt{1 - \frac{2\lambda^2}{\pi(\lambda^2+1)}}} \quad (5)$$

The goal then is to find the location(ξ) and skew (λ) parameters to satisfy the first two conditions. We use Nelder-Mead optimization [38] to minimize the sum of two squared distances: (1) the squared distance between the focal probability under the distribution X^* ($\mathbb{P}(X^* > x_{\text{FOCAL}})$) and the desired focal probability (p_{FOCAL}); and (2) the squared distance between the mode of the distribution X^* (mode(X^*)) and the original mode/mean (μ).

In practice, we are able to minimize this sum to ≈ 0; i.e., to find ξ and λ such that $\mathbb{P}(X^* > x_{\text{FOCAL}}) = p_{\text{FOCAL}}$ and mode$(X^*) = \mu$, satisfying conditions (1) and (2) above. An example of the resulting distribution is shown in Fig. 4b. It is similar to the Normal correction, except that its mode is the same as the mode of the original distribution. We provide code for this procedure in supplementary materials.

To apply these corrections to a decision-making task, we must know the intercept (α) and slope (β) parameters of the linear-in-probit function for a particular domain. It is also important to assess whether or not the theory of our bias corrections strategies holds up in practice. Thus, we require human subjects experiments.

4 EXPERIMENTS

We conduct two human subject experiments set in the context of U.S. Senate election forecasts. Here we assume people care about win probability and only correct for win (right-tailed) probabilities. **Experiment 1** estimates the intercept (α) and slope (β) parameters of the linear-in-probit function for subjective probabilities in the decision-making task of betting on election winners. These parameters allow us to derive the Normal and skew-Normal corrections described above (Sec. 3).

Experiment 2, which is preregistered, evaluates the two proposed bias corrections for win probabilities. We repeat the same procedure but show participants the bias-corrected distributions, expecting that the biases in participants' subjective probabilities of a candidate winning will decrease.

This section describes the experimental materials and design shared between the two experiments.

4.1 Materials

Cover story We use a cover story where participants read and interpret hypothetical U.S. Senate election forecasts. U.S. election forecasts have become controversial in recent years partly because people tend to misinterpret them [11, 58]. While media outlets such as FiveThirtyEight have been forecasting U.S. Senate elections since 2018, the general public is less familiar with U.S. Senate elections than a presidential election, which may reduce the effects of participants' prior knowledge on the experiments. We use the same cover story as Westwood et al. [58], but simplify it to focus on one candidate (Candidate A). As Westwood et al. only use text to convey a forecast in their experiments, we adjust the wording for histograms (bottom of Fig. 6).

Election forecasts State-of-the-art election forecasts estimate the percentage of the vote a candidate (or party) receiving, and convert the vote percentage distributions into the probability of one candidate (or party) winning the election [20, 32]. As the proposed corrections focus on Normal distributions, we generate election forecasts using Normal distributions with a standard deviation of 2.5% for vote percentage. These roughly correspond to the known standard deviation (2–4%) of senatorial polling in the U.S. [47]. We vary the means of the Normal distributions to generate different forecasts, and have ten values {46.80, 47.80, 48.53, 49.15, 49.72, 50.28, 50.85, 51.47, 52.20, 53.20}% for Candidate A's vote percentage. At a standard deviation of 2.5%, these mean values correspond to ten forecasts with various win probabilities: {0.1, 0.19, 0.28, 0.37, 0.46, 0.54, 0.63, 0.72, 0.81, 0.9}; i.e. the right-tailed probabilities $\mathbb{P}(X > 50\%) = 1 - F_{\text{Normal}}\big(50\%\ \mu, 2.5\%^2\big)$.

Text Similar to Westwood et al. [58], both experiments use a text representation to convey the probability of winning and vote percentage (Fig. 6 top). We explicitly tell participants the probability of Candidate A winning and a point prediction of the mode of the vote percentage distribution (the most likely outcome). A text description is also commonly used to convey election forecasts and polling results in media outlets [4, 50, 51].

Histogram The other representation in the experiments is a histogram of a forecast distribution, with the right-tailed probabilities $\mathbb{P}(X > 50\%)$ highlighted (Fig. 6 bottom). We use a violet color (#b150fb) for highlighting to avoid partisan effects. There are many visualizations for conveying forecast distributions (e.g., densities [12, 25], CDFs [12, 25], quantile dotplots [12, 27, 31], intervals [8, 27]). Because our focus is on adjusting the distribution, not the representation, we select one as a representative to assess our corrections. Histograms are a natural choice as they are commonly used in media outlets to convey election forecasts, especially senatorial forecasts [49–51].

Corrections After analyzing the data of Experiment 1, we would have known the intercepts (αs) and slopes (βs) of the linear-in-probit functions for text and histogram (detailed in Sec. 5). In Experiment 2, we derive both Normal and skew-Normal bias corrections for the original Normal forecast distributions. We recalculate the point predictions for text and regenerate histograms, preserving the same y-axis height and total area under the histogram as were in the first experiment.

4.2 Elicitation

For each of the ten forecasts (ten probabilities of Candidate A winning), we use three questions (Fig. 7). The first two ascertain whether participants can read the text (or histogram) and are used to examine the construct validity. The third elicitation induces participants to make decisions using probabilistic forecasts and is the focus of our analysis. It elicits the subjective probabilities people internalize and act on in their decision-making [36] and measures decision quality (see Sec. 2.1).

Elicitation 1: Likelihood. *On a scale from 0 (very unlikely) to 100 (very likely), how likely is Candidate A to win the election?* (Fig. 7a)

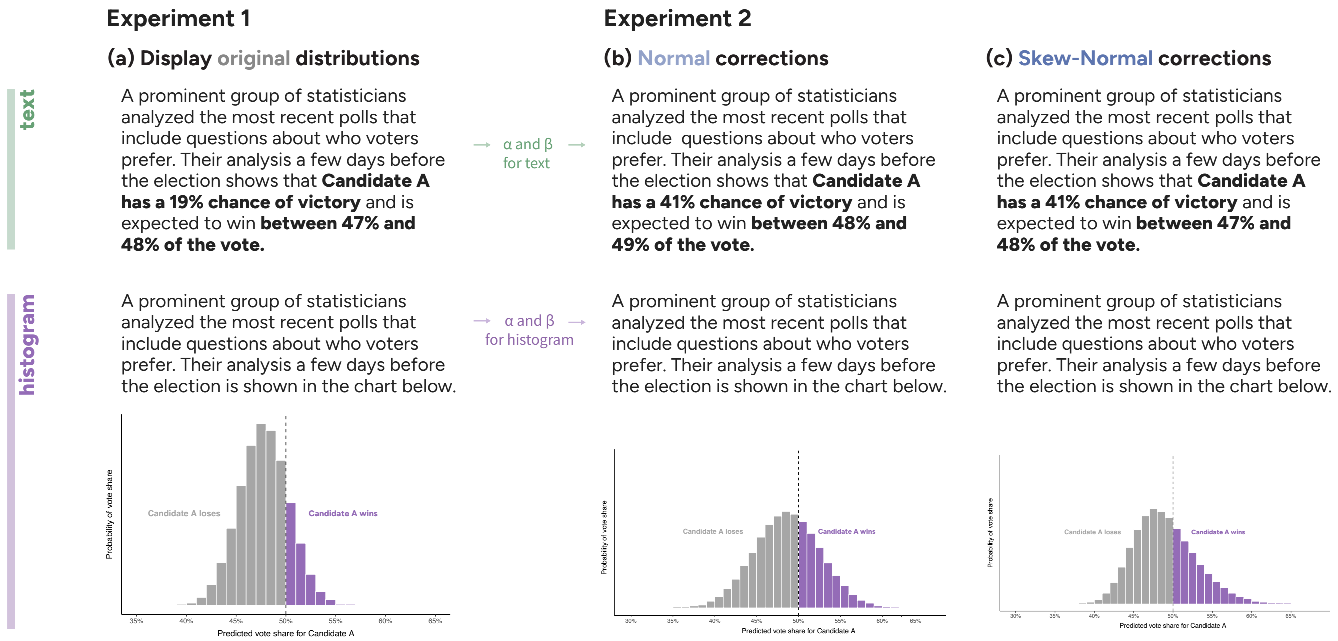


Figure 6: Examples of representations and bias-corrected forecast distributions.
Fig. (a) shows text (top) and the histogram (bottom) in Experiment 1. In this example forecast, the true probability of winning is 0.19, and the mode of the vote percentage distribution falls into 47%–48%.
Figs. (b) – (c) show the representations of the bias-corrected distributions in Experiment 2. For text, note that the differences in the bolded description. For histograms, the two corrections have the modes falling into 48%–49% and 47%–48%, respectively, and the areas highlighted are 0.37. Also, note that for histograms, we use a different color in this figure for aesthetic purposes.

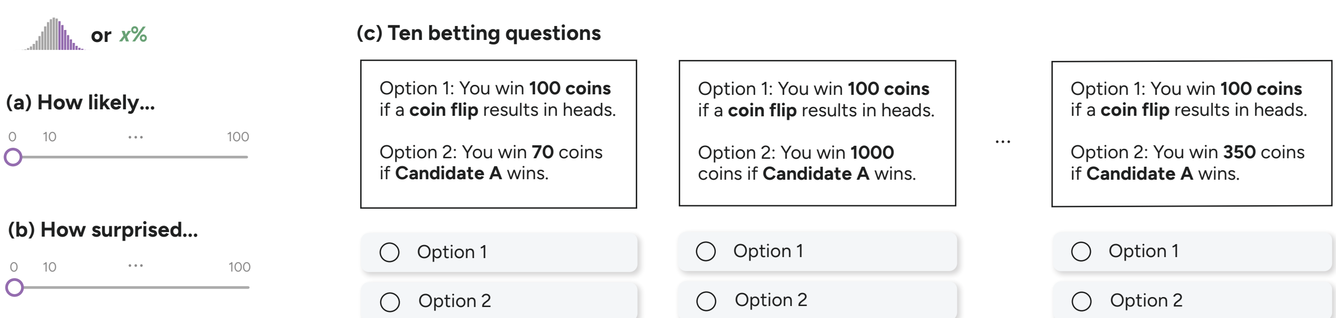


Figure 7: Illustration of the user interface for the three elicitations: For each of the ten forecasts, participants first answer (a) a likelihood question and (b) a surprise question. They then make (c) a sequence of 10 betting decisions, asking them to choose between two options. Each bet (a box) is presented separately, and the rewards in Option 2 are randomized.

Elicitation 2: Surprise. *On a scale from 0 (very unsurprised) to 100 (very surprised), how surprised would you be if Candidate A wins the election?* (Fig. 7b)

Elicitation 3: Betting. Participants are asked to make a sequence of ten binary decisions where they choose between two reward options (Fig. 7c):

> *Option 1: You win 100 coins if a coin flip results in heads.*
>
> *Option 2: You win {50, 60, 70, 80, 90, 100, 150, 200, 350, 1000} coins if Candidate A wins.*

As participants decide whether or not to take the bets, this elicitation invites a sequence of comparisons between 0.5 (the result of a coin flip) and participants' subjective probabilities of Candidate A winning. Comparisons to well-known frequency probabilities reliably elicit subjective probability in decision-making [40], especially under a betting task [3, 19, 26, 36]. Also, online prediction markets for U.S. elections have been active for decades [18], giving this task some real-world applicability. We refined the task by piloting several versions based on the literature, consulting with colleagues, fine-tuning the wording, and carefully checking quantitative and qualitative data from each pilot. We also incentivize participants, following the literature in subjective probability [19, 56], and simulate a winner for the election based on the forecast and inform participants that they will receive the coins as a study bonus (10,000 coins = $1 USD).

The two options are set as follows. The expected reward of Option 1 is 50 coins (100 coins · 0.5), and the expected reward of Option 2 is the reward times the true probability of winning ($reward \cdot p_{\text{TRUE}}$). Thus, the rewards in Option 2 are set to cover the range of values of $50/p_{\text{TRUE}}$ ($p_{\text{TRUE}} \in \{0.1...0.9\}$, see Sec. 4.1), the rewards at which a normative decision-maker would switch from Option 1 to Option 2. To avoid learning and order effects, the ten bets are presented one at a time in random order for each forecast.

4.3 Experimental design

Factorial design Both experiments follow mixed factorial designs. The within-subjects factor is the ten forecasts. The between-subjects factors are representations and corrections. Each participant sees ten forecasts, and the order is randomized. In Experiment 1, each participant is randomly assigned to either histogram or text. In Experiment 2, each participant is randomly assigned to one of the four combinations: {text, histogram} × {Normal correction, skew-Normal correction}. This design reliably measures subjective probabilities in decision-making and minimizes carryover and fatigue effects.

Training Both experiments include a training session for participants to ensure the construct validity of the study. The training session presents an example forecast and asks participants three questions: (1) which candidate is more likely to win; (2) what is Candidate A's chance of winning; and (3) which outcome is more likely. For histograms, we annotate the example histogram to explain the meaning of the highlighted (and gray) areas, which is the probability of Candidate A winning (or losing), and the meaning of the tallest bar, which is the most likely outcome (the mode of the distribution). The training session does not give feedback, and all participants in both experiments get at least one of the questions correct. We provide these in supplementary materials.

Procedure After participants enter the Qualtrics survey and consent, they first take part in the training session. Participants are then informed that the scenario and questions will always be the same, and they will receive a bonus of up to $2.30 USD based on their responses. They then finish ten forecasts, and in each forecast, they answer the likelihood and surprise questions, and make ten betting choices. After ten forecasts, they are asked for their strategies in the questions and additional feedback. Each participant answers 10 likelihood and 10 surprise questions and makes $10 \cdot 10 = 100$ binary betting choices, taking about 15 minutes (see the details in Appx. D).

Participants We recruit all participants from `Prolific.co` and limit experiments to desktop users, as forecast websites are often designed for desktop use. Because of the U.S. election context, we use U.S. demographically-representative samples provided by `Prolific.co`. In Experiment 1, we request a minimal sample size of 300, as we do not have an expectation of effect size; each of the two conditions has about 150 participants. In Experiment 2, we have four conditions. As the precision of our estimates from Experiment 1 is satisfactory, we request the same per-condition sample size (150) to ensure similar precision, resulting in a request of 600 participants.[7] The exact number of participants depends on `Prolific.co`'s sampling strategies, and we obtain 306[8] and 603 participants for the two experiments (see demographics and breakdowns in Appx. D). The pilot and previous participants are excluded from later experiments. The study was approved by the Institutional Review Board (IRB) at Northwestern University as exempt human subjects research (STU00215415).

Compensation We pay each participant $4.00 USD for completing an experiment. For each forecast, we simulate the winner using a random number generator and pay them the reward based on their responses. The means of resulting bonuses are $0.98 ($\sigma = 0.23$) and $1.02 ($\sigma = 0.24$) USD for the two experiments.

5 MODELING SUBJECTIVE PROBABILITIES

With participants' binary responses to the betting questions, we use a nonlinear Bayesian multilevel model to infer participants' subjective probabilities of Candidate A winning for this decision-making task.

5.1 Model specification

As described in Sec. 3.2, we model subjective probabilities as a linear-in-probit function of the true probabilities. Because participants were asked to choose between a coin flip and a bet of Candidate A winning, we presume they make decisions based on the following rule: they are more likely to take the bet ($p_{\text{TAKEBET}} > 0.5$) if their subjective expected reward is greater than the expected reward of a coin flip ($reward \cdot p_{\text{SUBJECTIVE}} > 50$ coins). Here, we use a scaling factor θ that determines how sensitive people are to differences in rewards, and derive a model formula that satisfies this requirement:

$$
\begin{aligned}
\text{subjective}\mathbb{E}(\textit{taking the bet}) &> \mathbb{E}(\textit{flipping the coin}) \\
\textit{reward} \cdot p_{\text{SUBJECTIVE}} &> 0.5 \cdot 100 \\
\textit{reward} \cdot \text{lpr}(p_{\text{TRUE}}) &> 50 \\
\textit{reward} \cdot \text{probit}^{-1}(\alpha + \beta \cdot \text{probit}(p_{\text{TRUE}})) &> 50 \\
\textit{reward} \cdot \text{probit}^{-1}(\alpha + \beta \cdot \text{probit}(p_{\text{TRUE}})) - 50 &> 0 \\
\theta \cdot \left[\textit{reward} \cdot \text{probit}^{-1}(\alpha + \beta \cdot \text{probit}(p_{\text{TRUE}})) - 50\right] &> 0 \qquad \theta > 0 \\
\text{logit}(p_{\text{TAKEBET}}) = \theta \cdot \left[\textit{reward} \cdot \text{probit}^{-1}(\alpha + \beta \cdot \text{probit}(p_{\text{TRUE}})) - 50\right] &> \text{logit}(0.5) \qquad \theta > 0
\end{aligned}
$$

Built around this core formula representing the decision rule for the task, the full model formula is:

`line 1` $$takebet \sim \text{Bernoulli}(p_{\text{TAKEBET}})$$

`line 2` $$\text{logit}(p_{\text{TAKEBET}}) = \underbrace{\theta_r}_{\text{a scaling factor}} \cdot \left[\overbrace{\left[\underbrace{reward}_{\text{one bet reward}} \cdot \overbrace{\text{probit}^{-1}(\underbrace{\alpha_{rj}}_{\text{intercept}} + \underbrace{\beta_{rj}}_{\text{slope}} \cdot \text{probit}(\underbrace{p_{\text{TRUE}}}_{\text{the true probability of winning}}))}^{\text{the subjective probability of Candidate A winning}} \right]}^{\text{the subjective expected reward of one bet}} - \overbrace{50}^{\text{the expected reward of a coin flip}} \right]$$

`line 3` $$\alpha_{rj} = \alpha_r + \delta_{\alpha j}$$

`line 4` $$\beta_{rj} = \beta_r + \delta_{\beta j}$$

`line 5` $$\begin{bmatrix} \delta_{\alpha j} \\ \delta_{\beta j} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right)$$

`line 6` $$\theta_r \sim \text{Exponential}(1)$$

`line 7` $$r \in \{1...2\} \qquad \textit{(two representations)}$$

`line 8` $$j \in \{1...J\} \qquad \textit{(J participants)}$$

`line 1` We model participants' decisions as a Bernoulli distribution, with the probability of taking a bet (p_{TAKEBET}) as a function of their decision rule described above.

`line 2` In the logit space, this line is the participants' decision rule: when $reward \cdot p_{\text{SUBJECTIVE}} > 50$, it ensures $p_{\text{TAKEBET}} > 0.5$. In other words, participants are more likely to take the bet than not if their subjective expected reward is greater than the expected reward of a coin flip. Within `line 2` , blue indicates model parameters to be estimated, orange indicates input to the model (predictors), and gray indicates transformed parameters and constants.

`lines 3–5` We expect that both the linear-in-probit intercepts (αs) and slopes (βs) vary with different representations and participants. α_{rj} and β_{rj} are the intercept and slope for participant j with representation r. α_r and β_r (without participant j) are the intercept and slope for an *average* participant ($\delta_{\alpha j} = 0$, $\delta_{\beta j} = 0$) with representation r. The posterior medians of this average participant will be used to construct corrected distributions. Because different participants may have personal strategies, we model participant-level slopes and intercepts as random effects. The $\delta_{\alpha j}$ and $\delta_{\beta j}$ capture the differences between each participant's own intercepts and slopes compared to the average participant's for each representation.

`line 6` The non-negative scaling factor θ may also vary in different representations. As it is a nuisance parameter and we are interested only in α and β, for simplicity, we do not model participant-level differences in θ.

[7] We used our Experiment 1 data to simulate Experiment 2 results at various sample sizes, and concluded that using the same overall sample size would yield much less precise estimates.

[8] The analysis of Experiment 1 includes six participants who did not finish the experiment but at least one forecast. Because we preregistered this analysis and used it in Experiment 2, we include these participants in our report. We provide the analysis excluding these participants in supplementary materials, which yields almost identical results. The analysis of Experiment 2 includes only participants who accomplished the experiment.

`priors` In the logit space, we center priors for α_{rj} at 0 and β_{rj} at 1, indicating no bias in subjective probability. We allow them to approach the extreme distorted cases and thus have $\alpha_{rj} \sim \text{Normal}(0, 1)$ and $\beta_{rj} \sim \text{Normal}(1, 2)$ (see Fig. 3 and Appx. A). For covariance, we let $\Sigma = \text{diag}(\tau)\Omega\text{diag}(\tau)$, then τ is a vector of standard deviations of $\delta_{\alpha j}$ and $\delta_{\beta j}$, and Ω is their correlation matrix. We expect some variance in slopes and intercepts and have $\tau \sim \text{half-Normal}(0, 0.5)$ as priors. We also expect a weak correlation between participants' slopes and intercepts and set a $\Omega \sim \text{LKJcorr}(2)$ prior.

Experiment 2 We expect that the two corrections have different effects on participants' subjective probabilities in this decision-making task. Therefore, the model for Experiment 2 replaces representations with the interaction between representations and corrections in all lines, i.e., $r \in \{1...4\}$ from the four combinations: {text, histogram} × {Normal correction, skew-Normal correction}. The priors and other terms are otherwise the same. We preregistered the model specification and priors for Experiment 2.

Implementation We implemented these models using `R 4.2.0`, `Rstan 2.21.5` [53], `CmdStanR 0.5.2` [14], `brms 2.17.0` [5], and `tidybayes 3.0.2` [29]. We use the logit approximation of probit [1], which is $\text{logit}^{-1}(1.7 \cdot \alpha + \beta x) = \text{probit}^{-1}(\alpha + \beta x)$, to help the model converge. We inspected the minimal bulk and tail effective sample size (ESS) to ensure reliable estimates; and they are 1613 and 2693, both coming from the average participant's β for histograms in Experiment 1. We also examined $\hat{R}$ values (1.0) to ensure model convergence. We provide code and fitted model files in supplementary materials (`*.Rmd` and `*.rds` files).

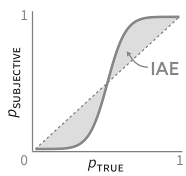
5.2 Derived measures

We derive three measures from participants' subjective probabilities of Candidate A winning. The first two are part of the linear-in-probit subjective probability function. The third measure describes the overall deviation from the true probabilities.

Intercept (α) This is the estimated intercept of the linear-in-probit function. If participants do not systematically under- or over-estimate the 0.5 probability, the intercept should be close to 0. If the intercept deviates from 0, participants believe a different probability is 0.5.

Slope (β) This is the estimated slope of the linear-in-probit function. If participants' subjective probabilities are not distorted, the slope should be close to 1. If the slope is larger than 1, participants may systematically underestimate small probabilities and overestimate large probabilities (the intercept determines the threshold).

Integrated absolute error (IAE) The last measure combines both the intercept and slope to provide an overall estimate of decision quality. This measure integrates the difference between subjective and true probabilities in the range of 0 and 1, defined as $\int_0^1 |p_{\text{SUBJECTIVE}} - p_{\text{TRUE}}| dp_{\text{TRUE}}$. It can be interpreted as the average bias in subjective probabilities. Visually, this measure is the area between the linear-in-probit function and the diagonal line $y = x$, as shown in the figure on the right side.



Because of our use of a Bayesian modeling approach, the uncertainty in these measures is quantified by posterior probability distributions from the models. Similar to our corrections, we only use the posterior estimates conditional on the average participant (i.e., setting participants' random effects to zero) to calculate these measures. The exploration of the scaling factor θ is provided as Appx. G and in supplementary materials.

6 RESULTS

With the models and measures, for each experiment, we first present the modeled subjective probabilities of winning for an average participant (Figs. 8a and 9a). We then present posterior distributions of the three measures as well as comparisons between the two experiments (Figs. 8b-d and 9b-d). We also report the standard deviations of random effects in text and provide an exploration of participant-level random effects in Appx. F; those results lead to similar conclusions but are more difficult to interpret due to the non-linearity. All the results are reported as medians and 95% quantile credible intervals (CIs; Bayesian analog to confidence intervals). The analyses of likelihood and surprise questions are provided in Appx. E to ensure the construct validity.

We also conducted two post hoc analyses: (1) calculating the total expected rewards as a post hoc alternative measure for decision quality and (2) coding participants' self-reported decision strategies based on their free-text responses.

6.1 Experiment 1: Decision-making with the original forecast distributions

Subjective probability Participants' **subjective probabilities** are biased (Fig. 8a) when they make decisions based on the election forecasts in both text and histogram representations. They underestimate small probabilities and overestimate large probabilities. With text, their subjective probabilities have a more distorted S-shape. With histograms, their subjective probabilities are less biased for large probabilities, e.g., $p_{\text{TRUE}} > .8$: the S-shape is closer to the diagonal line.

Measures The linear-in-probit **intercepts** (αs; Fig. 8b) for text and histograms are similar, and they both deviate from 0 (-0.34 [-0.44, -0.24]), indicating a non-zero fixed point in the probit space. The linear-in-probit **slopes** (βs; Fig. 8c) also deviate from 1, but histograms yield smaller deviations (1.58 [1.31, 1.86]) than text (2.44 [2.16, 2.74]). Combining them, the **integrated absolute errors** (IAE; Fig. 8d) show that both text and histograms lead to substantial biases in average participants' subjective probabilities (0.092 [0.074, 0.11] and 0.13 [0.12, 0.14]). In this decision-making task, average participants' subjective probabilities, on average, are about 10 percentage points away from the true probabilities. The standard deviations of random intercepts for α and β are 0.62 [0.56, 0.68] and 1.67 [1.49, 1.86], respectively.

Summary Together, the results of this experiment suggest that there are systematic biases in participants' subjective probabilities in making decisions from probabilistic election forecasts, regardless of the representation we chose.

The results of Experiment 1 also give us the linear-in-probit intercepts and slopes for this decision-making task to derive the corrected distributions. Because our goal is to create a single corrected forecast, i.e., not to tailor forecast distributions to each participant, our bias corrections for Experiment 2 use the median intercept and slope from posterior estimates conditional on an average participant. In principle, we could use this model to derive participant-level corrections; however, this level of tailoring would be difficult to accomplish in a journalistic setting like election forecasting. Here we leave participant-level corrections for discussion and future work (see Sec. 7.1). Thus, for text, we use the posterior medians $\alpha = -0.34$ and $\beta = 2.44$ to derive bias-corrected distributions for Experiment 2; for histogram, we use the posterior medians $\alpha = -0.34$ and $\beta = 1.58$ to derive bias-corrected distributions.

6.2 Experiment 2: Decision-making with the bias-corrected forecast distributions

In Experiment 2, we show participants the bias-corrected distributions in text or as histograms and repeat the same procedure. We expect that these corrections will improve participants' subjective probabilities in decision-making, reducing biases and bringing them closer to the true probabilities. We preregister three measures: the linear-in-probit (1) intercept and (2) slope, as well as (3) integrated absolute error.

Subjective probability Visually, participants' subjective probabilities look much closer to the true probabilities across all conditions, regardless of the representation or correction we chose for the experiments (Fig. 9a). In particular, participants improve their underestimation of small probabilities, although it appears that both Normal and skew-Normal corrections slightly over-correct large probabilities, making large subjective probabilities slightly further deviate from the true probabilities.

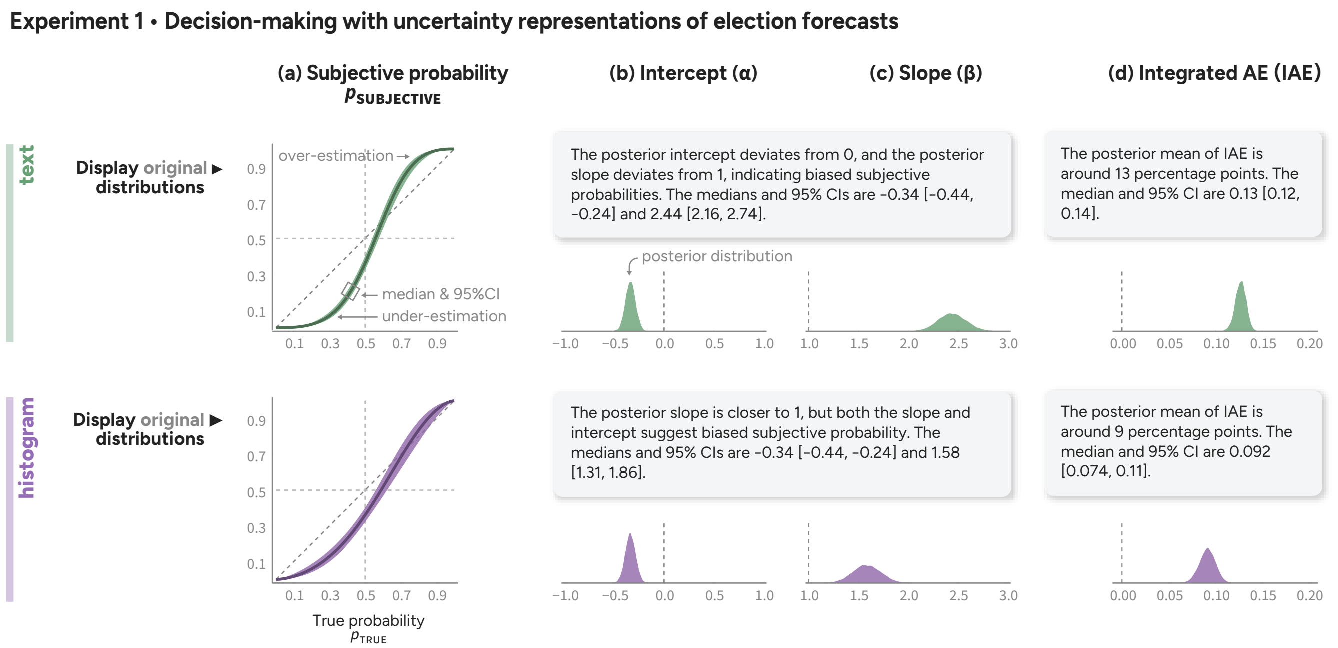


Figure 8: The main results of Experiment 1. All are posterior estimates from the model. (a) We find substantial biases in participants' subjective probabilities. These are evidenced by (b) the intercepts (αs) of the linear-in-probit functions deviating from 0 and (c) the slopes (βs) deviating from 1. (d) The integrated absolute errors between subjective and true probabilities are 0.13 [0.12, 0.14] (text) and 0.092 [0.074, 0.11] (histogram), indicating about 10 percentage points of biases.

Preregistered measures Both Normal and skew-Normal corrections debias the linear-in-probit **intercepts** (αs; Fig. 9b) for text and histograms, bringing them closer to 0. In particular, the skew-Normal correction brings the intercepts very close to 0, e.g., -0.032 [-0.14, 0.081] for histograms; the Normal correction slightly overcorrects the intercepts, e.g., 0.15 [0.032, 0.27] for histograms. The standard deviation of random intercepts for α is 0.69 [0.64, 0.74].

Both Normal and skew-Normal corrections debias the **slopes** (βs; Fig. 9c) for text and bring them much closer to 1, e.g., 1.43 [1.21, 1.66]; neither correction debiases the slopes for histograms (though this was already around 1.5 in Experiment 1). The standard deviation of random intercepts for β is 1.32 [1.22, 1.43].

Combining them, both corrections reduce the **integrated absolute errors** (IAE; Fig. 9d) and improve decision quality from both representations, and the improvement for text is very substantial, from more than 10 percentage points to about 5 percentage points (i.e., reducing 50% of the biases), suggesting a large improvement. Between the two corrections, the skew-Normal correction for text makes participants slightly less biased in their decision-making, but these two corrections are similar for histograms.

Summary Both corrections improve participants' subjective probabilities for text and histograms; they also bring the subjective probabilities of the two different representations closer to each other. These corrections have a bigger impact on text because of the larger innate biases found in Experiment 1. Between the two corrections, the skew-Normal correction may be slightly more effective (perhaps due to its preserving of the mode of a forecast distribution). But the shift in the mode is subtle (Fig. 5), which may explain why the differences between the two corrections are small.

6.3 Post hoc analysis: Corroborating the improvement in decision quality

Method Because of our bonus mechanism (see Sec. 4.1), expected rewards can also be used to measure participants' decision quality. Given the posterior probability of (i.e., setting participants' random effects to zero) taking a bet, we can calculate the expected reward of an average participant's decision on that bet:

$$p_{\text{TAKEBET}} \cdot (p_{\text{TRUE}} \cdot \mathit{reward}) + (1 - p_{\text{TAKEBET}}) \cdot 50 \qquad (6)$$

We accumulate the average participant's expected reward for all 100 bets and report the total expected rewards in Fig. 10. Unlike the results above, this measure is not preregistered; we use it as a reasonability check.

Results Both corrections result in a substantial improvement in the total expected rewards for the average participant, from 11.31k [11.42k, 11.57k] to 11.63k [11.61k, 11.65k] coins depending on representation and correction, although the absolute improvement (about 300 coins) is small compared to the total expected reward under an optimal strategy (11,879 coins). These results are similar to those of integrated absolute errors of subjective probability. This is because participants' subjective probabilities underlie their decisions, and the optimal decision is achieved when participants' subjective probabilities match the true probabilities.

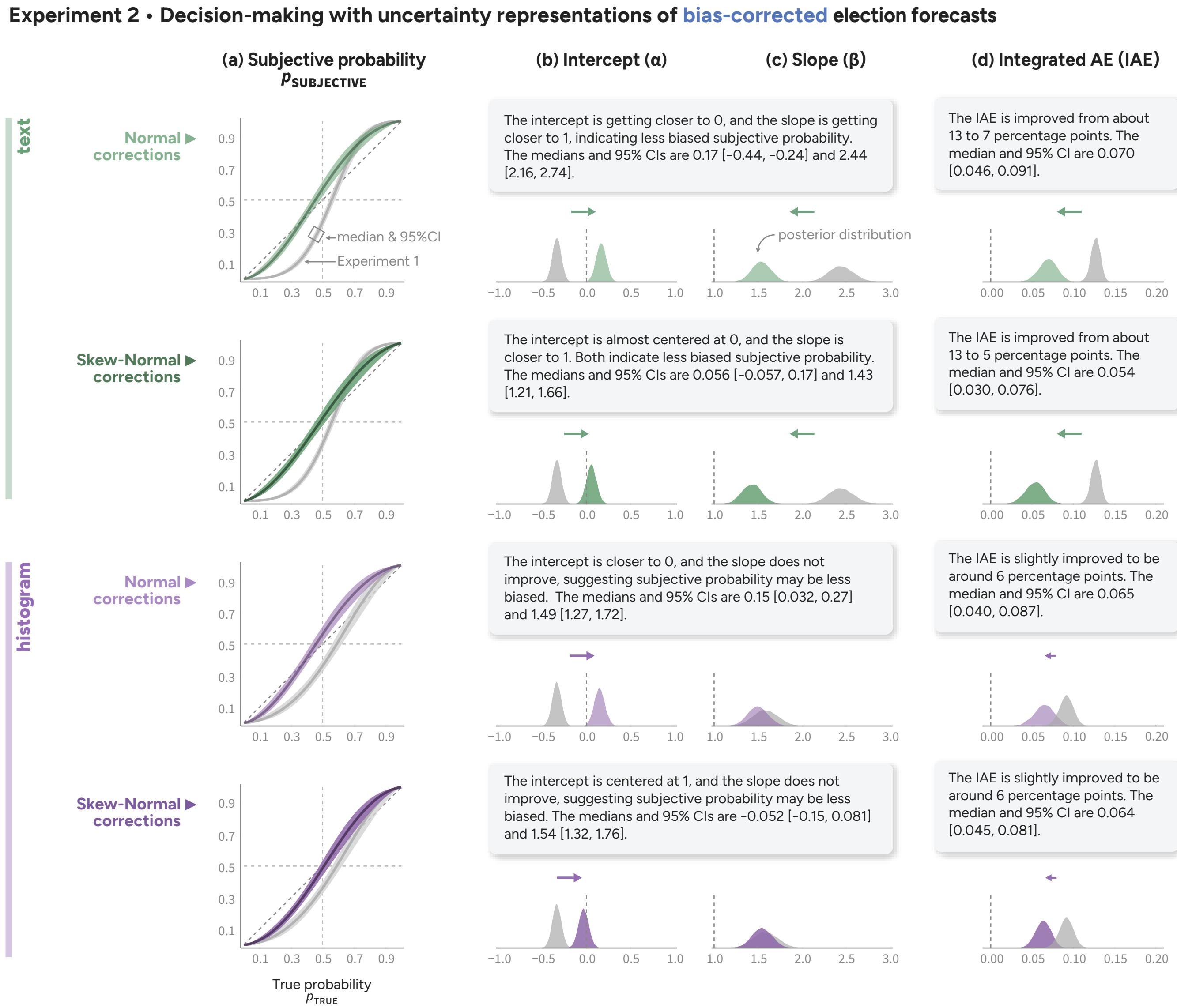


Figure 9: The main results of Experiment 2. All are posterior estimates from the model. (a) When showing bias-corrected distributions to participants, visually, we find the biases in participants' subjective probabilities decrease. These are evidenced by (b) the intercepts (αs) of the linear-in-probit functions are closer to 0, and (c) the slopes (βs) for text are closer to 1. (d) The integrated absolute errors between subjective and true probabilities are also reduced for text (from 0.13 [0.12, 0.14] to 0.054 [0.030, 0.076]) and slightly for histograms.

6.4 Post hoc analysis: Coding decision strategies

We asked participants to report their strategies as free-text responses and performed qualitative coding to gain further insight into this decision-making task.

Method Because we had over 900 responses, one of us (the primary coder) looked through comprehensible responses and categorized them until they exhausted the types of participants' strategies. All the authors then discussed and refined the coding scheme as presented by the primary coder using representative examples. This proceeded in several rounds until all were satisfied with the code book. The primary coder then randomly sampled 200 responses (100 from histogram and 100 from text) and coded them to estimate the prevalence of each strategy. Of these, 33 were too vague to identify a clear strategy and were removed from the analysis. We coded the remaining 167 responses based on whether the participant used (1) Candidate A's win probability, (2) the reward for each option,

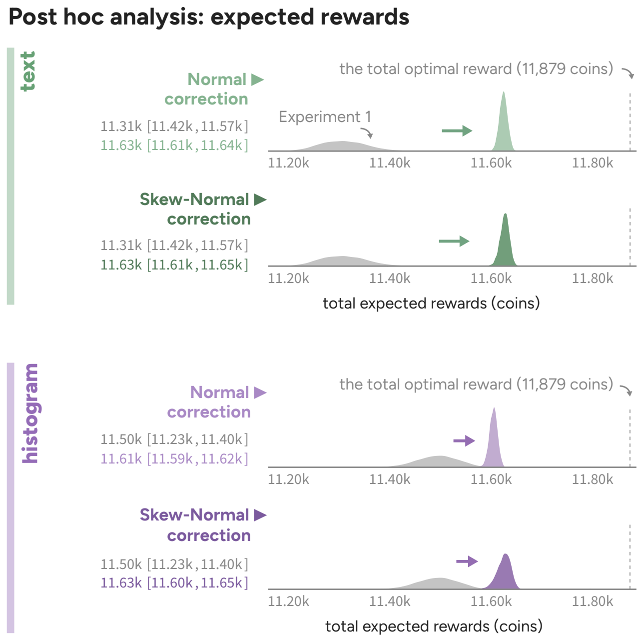


Figure 10: The two corrections also improve the total expected rewards for an average participant. This alternative measure of decision quality corroborates our preregistered measure, IAE of subjective probability.

(3) both, or (4) neither. We also cataloged specific strategies under each category, reporting the results in Table 1.

Results Around 42% of participants used both the win probability and the reward in betting. This included participants who calculated expected values for each option, participants who described making some kind of tradeoff between win probability and reward, and participants who described using probability and reward, but did not outline a more specific strategy. Around 47% of participants described various strategies using only the win probability, whether this was by always selecting the option with the highest win probability, or through some other probability-based decision rule. Around 9% of participants only used the reward in their decision-making, by always selecting the option with the highest reward. Around 2% of participants did not refer to probability or reward at all.

It is notable that 42% of participants used both probability and rewards; i.e., the information that forms the basis of our presumptive decision rule: expected value. They might not all have calculated an expected value, but it is reasonable to think that many of their strategies could approximate this rule. While others might have not followed an approximation, the improvements in subjective probability suggest our corrections have some robustness to variance in strategies. However, around 11% of participants did not describe using probability at all. They were likely to be unaffected by our corrections. Similarly, some participants might not have interpreted the task correctly (e.g., "I chose the coin flip because it has a greater chance of winning"). Both help explain why our corrections are imperfect.

7 DISCUSSION

7.1 Why is the correction imperfect?

While our correction methods improve subjective probability and overall decision quality, it is worth interrogating why the slopes of the linear-in-probit function were not completely debiased. One explanation is that participants may be using multiple, different strategies: Kale et al. [27], for example, found evidence that different people employ different strategies when attempting to make decisions from uncertainty visualizations, some of which are better matched to the decision task than others. Our qualitative results revealed a similar variety of strategies and heuristics, only about half of which correspond approximately to our decision rule (and even then, many of these not precisely). Other heuristics may not respond to changes in the linear-in-probit parameters in the way our model expects: e.g., people using a heuristic of always picking the highest reward may not change their decisions at all, even if the model predicts that someone with their particular linear-in-probit curve should change. Other cognitive biases (e.g., preferring an immediate reward [45]) may also affect decisions in ways not captured by the model (a coin flip may sound more tangible than a hypothetical election—and we did see some participants always take the coin flip).[9] Another ostensible explanation may be that individuals focus on different visualization properties than the ones we corrected for (e.g., left-tailed lose probabilities instead of right-tailed win probabilities); while we did not see explicit evidence for this (participants talked mostly about the reward or the win probability, not the lose probability), it may be worth investigating more directly, perhaps via eye- or mouse-tracking studies. In any case, it is clear that some individuals' strategies will not be impacted by our corrections as we expected.

Whatever the cause, imperfect subjective probability correction is a natural result of using a model that simplifies complex human behavior to a two-parameter equation. It is notable that even this simplified model is able to produce an effective and robust correction. Future work could attempt to model the mixture of strategies employed within a population to develop more precise corrections, or even develop personalized corrections (this could help address the variance in IAE across participants; see Appx. F). Our work sets a baseline of comparison against which more complicated correction methods can be judged.

7.2 Applying corrections in practice

It is exciting to see that both correction methods improve decision quality for both representations. After correction, both representations elicit very similar performance, and the impact of the correction on error (reduction on the order of 5 percentage points) is large for uncertainty visualization. If this result holds across other representations, for tasks where it is possible to apply such corrections, the particular representation used may matter less. This has important implications when some representations are harder for some people to understand than others, e.g., if working memory

[9] It may be tempting to look to perceptual biases for an explanation as well; e.g., tendencies to underestimate areas according to Stevens' power law [54]. We think this is a less likely explanation, as the mathematical basis of the linear-in-log-odds model (and therefore the linear-in-probit approximation) in Stevens' power law means such biases should be accommodated by changes in the parameters of the linear-in-probit function; our secondary examination of likelihood results (Appx. E) corroborates this.

Table 1: Participants' betting strategies.

Category	Specific strategy	Examples	Est. %
Probability & reward	Using expected values	"Compared the EV of Option 2 (probability X potential payoff) with EV of Option 1 (50)" "I multiplied the coins won by the projected chance of victory"	6%
Probability & reward	A tradeoff between probability and reward	"I chose what I thought were the better odds. Except I always picked 1000 because why not?" "I would always hit option 2 if I thought A was going to win or if it was 350+ to win"	28%
Probability & reward	Unspecified strategy using probability and reward	"Based upon probability and amount of gamble" "I tried to pick the bigger coin amount that had more likelihood of winning"	8%
Probability	Using win probability	"I would choose the 'safer bet' " "What I thought had the best odds" "If the candidate had better than 70%"	47%
Reward	Choosing the highest reward	"I picked the larger coin amount no matter by who" "Everyone wants the higher amount so I took that just in case a miracle happened and they did win"	9%
Neither	Choosing the coin flip	"I wanted a 50/50 chance of winning."	1%
Neither	Choosing at random	"I randomly picked" "Kind of back and forth to make the odd go either way."	2%

capacity limits some people's ability to take advantage of some representations [6].

That said, there are practical issues with applying these corrections. First, the decision task and associated heuristics must be unambiguous; in election forecasting, we have assumed people are interested in right-tailed (win) probabilities, not left-tailed probabilities—and this is largely corroborated by our qualitative results. If a task called for people to focus on other aspects of a distribution (e.g., its variance), a different correction would be needed. In the election forecasting scenario, because the intercept (α) is nonzero, a slightly different correction would be applied depending on if the viewer is interested in whether Candidate A wins (right-tailed probability) or loses (left-tailed). One compromise would be to pick a correction that is symmetric, even if imperfect: since α in this task is relatively small, we could fix it to 0 so that the corrections for left- and right-tailed probabilities are equivalent. This would amount to multiplying the forecast distribution standard deviation by $\beta \approx 2$ (the approximate value of β estimated in Experiment 1); and this would essentially attempt to correct for people's tendency to "round to 100%" or "round to 0%" (see Fig. 3a right column, where $\beta = 2$). In this way, knowledge of domain tasks and the biases at play may be used to construct a correction tailored to a particular use case; work on guidelines for applying distribution corrections in practice may therefore be fruitful.

Second, distributions other than the Normal may require tailored corrections, depending on the task. We provide the closed-form Normal correction because Normal distributions are ubiquitous in uncertainty quantification, and this simpler correction may be more accessible to practitioners. For other distributions, if the goal is to correct the CDF or CCDF, it is not necessary to know a closed-form correction nor to tailor the correction to the distribution. The generic correction formula in Sec. 3.3 can be applied directly to either the parametric CDF or CCDF of a distribution, or an approximate CDF or CCDF estimated from a sample (see Appx. H). However, if domain-specific concerns complicate the task—like the mode crossing the 50% line in an election forecast—more tailored corrections like the skew-Normal correction may be needed.

Third, it is necessary to elicit the parameters of any correction before applying it in practice. At the very least, there are several domains of sufficient societal importance that this elicitation exercise is worthwhile: e.g., election forecasts, climate change forecasts, and epidemiological modeling (as in the COVID-19 pandemic and its associated forecasts). For journalists working in U.S. election forecasting, corrections using our parameter estimates should be appropriate, as we used U.S. demographically-balanced samples. Future research could use our methods to estimate correction parameters for other domains, which could then be adopted by practitioners. Such efforts would lead to a clearer picture of how and why subjective probability varies across domains, yielding an improved understanding of people's decision-making under uncertainty.

7.3 Ethics of subjective probability correction

One possible objection to applying these corrections may be that adjusting probabilities is not transparent (or worse, amounts to lying). We believe this question is not so simple, and rests somewhat on a foundational question of uncertainty communication: is the goal to communicate mathematically precise probabilities (assuming a forecaster's probabilities are "true"—which is already dubious), or is the goal to *induce* reasonable subjective probabilities in the viewer? If a viewer's decision-making process is better aligned to forecasts that overstate uncertainty, e.g., by multiplying the standard deviation by a factor of ≈ 2, is applying this correction unethical? Or if it is ethical to correct color scales to be perceived as uniform even when mathematically they are not, is it also ethical to correct probabilities to be acted on more normatively even if those probabilities are not displayed exactly as calculated by a forecast? And if the answers to these questions differ: why?

We do not claim to have perfect answers to these questions, and suspect that the answers will vary by decision context, audience values, etc. Some forecasters appear already to have answered "yes": weather forecasters, for example, may over-predict the probability of rain, because their audience is happier when a forecast for rain is wrong (they don't get wet) than when a forecast for no rain is wrong (and they get wet) [48]. An understanding of the stakes in a decision, the expertise of the decision-maker, and the communication norms of a domain should determine whether and how bias corrections might be applied.

These questions do suggest a need for further study and guidance in how to report forecasts if a correction is applied. One approach may be to apply such corrections transparently: e.g., to experiment with showing an uncorrected and corrected distribution together, or to include descriptions of the correction and its rationale in text even if the uncorrected distribution is not shown. The impact of such approaches on decision-making is worth studying: if people are told the correction is applied to aid their decision-making, does this negate the benefit of the correction?

Another approach may be to adopt communication strategies that have the effect of a bias correction without applying it to the distribution representation itself. For example, Padilla et al. [42] found that qualitative expressions of low forecaster confidence, e.g., labeling a forecast "low confidence", paired with a forecast distribution, had essentially the equivalent effect of increasing the standard deviation. If a qualitative expression could be found that approximately aligns with the desired increase in standard deviation implied by the slope (β) for a domain, this may be an alternative to changing the representation itself. Overall, we believe that careful consideration of the values of the audience, the decision-making context, and further study of approaches to bias correction should allow visualization designers to more effectively—and ethically—construct uncertainty representations.

8 CONCLUSION

We propose a new approach to improve uncertainty communication: we can fix uncertainty representations but change the distribution being displayed to improve people's decision quality. Our approach corrects biases in subjective probabilities for uncertainty representations based on models of people's beliefs. We derive two corrections tailored for Normal distributions and show how to estimate the parameters for these corrections empirically. We also demonstrate that the corrections reduce biases in people's subjective probabilities and improve their decision quality. Our approach can be applied to any visual representation where the subjective probability function is known, and it can be generalized to any univariate probability or confidence distribution, giving it broad applicability. That said, questions remain about how to tailor the correction to the domain tasks (and concordant biases), and how to transparently apply such corrections in practice. Overall, our work opens a new avenue for subjective probability correction in uncertainty communication, providing a promising tool for decision-making under uncertainty.

9 AUTHOR CONTRIBUTIONS

All three authors contributed to experimental design and manuscript editing. Fumeng Yang conducted the quantitative analyses, prepared the manuscript, and led the submitting process. Maryam Hedayati implemented the experiments, collected the data, and conducted the qualitative analysis. Matthew Kay conceptualized and supervised the work, and prepared the manuscript.

ACKNOWLEDGMENTS

This research was supported by NSF IIS-1910431 and by NSF 2127309 to the Computing Research Association for the CIFellows Project. The authors thank the anonymous reviewers for their insightful comments. The authors also thank Alex Kale for his help with the experimental design, and Abhraneel Sarma and Hyeok Kim for their valuable feedback on the manuscript.

REFERENCES

[1] Takeshi Amemiya. 1981. Qualitative response models: A survey. *Journal of economic literature* 19, 4 (1981), 1483–1536.
[2] Francis J. Anscombe, Robert J. Aumann. 1963. A definition of subjective probability. *Annals of mathematical statistics* 34, 1 (1963), 199–205.
[3] Melanie Bancilhon, Zhengliang Liu, and Alvitta Ottley. 2020. Let's Gamble: How a Poor Visualization Can Elicit Risky Behavior. In *2020 IEEE Visualization Conference (VIS), short paper*. 196–200. https://doi.org/10.1109/VIS47514.2020.00046
[4] Ryan Best et al. 2022. Latest Polls. https://projects.fivethirtyeight.com/polls/
[5] Paul-Christian Bürkner et al. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80, 1 (2017), 1–28. https://doi.org/10.18637/jss.v080.i01
[6] Spencer C Castro, P Samuel Quinan, Helia Hosseinpour, and Lace Padilla. 2021. Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx. *IEEE transactions on visualization and computer graphics* 28, 1 (2021), 411–421.
[7] Roger Cooke et al. 1991. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press on Demand.
[8] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2142–2151.
[9] Michael Correll and Jeffery Heer. 2017. Surprise! Bayesian Weighting for De-Biasing Thematic Maps. *IEEE Transactions on Visualization & Computer Graphics* 23, 01 (jan 2017), 651–660. https://doi.org/10.1109/TVCG.2016.2598618
[10] Michael Correll, Dominik Moritz, and Jeffrey Heer. 2018. Value-Suppressing Uncertainty Palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174216
[11] Nicholas Diakopoulos. 2022. Predictive Journalism: On the Role of Computational Prospection in News Media. *Tow Center for Digital Journalism* (2022). https://doi.org/10.2139/ssrn.4092033
[12] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. 1–12. https://doi.org/10.1145/3173574.3173718
[13] James John Flannery. 1971. The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *Cartographica: The International Journal for Geographic Information and Geovisualization* 8, 2 (1971), 96–109.
[14] Jonah Gabry and Rok Češnovar. 2020. CmdStanR: the R interface to CmdStan. https://mc-stan.org/users/interfaces/cmdstan
[15] Charles R Gallistel, Monika Krishan, Ye Liu, Reilly Miller, and Peter E Latham. 2014. The perception of probability. *Psychological Review* 121, 1 (2014), 96.
[16] Benny Geys. 2006. 'Rational'theories of voter turnout: a review. *Political Studies Review* 4, 1 (2006), 16–35.
[17] Gerd Gigerenzer. 1996. The psychology of good judgment: frequency formats and simple algorithms. *Medical decision making* 16, 3 (1996), 273–280.
[18] John W. Goodell, Richard J. McGee, and Frank McGroarty. 2020. Election uncertainty, economic policy uncertainty and financial market uncertainty: A prediction market analysis. *Journal of Banking & Finance* 110 (2020), 105684. https://doi.org/10.1016/j.jbankfin.2019.105684
[19] J. M. Hampton, P. G. Moore, and H. Thomas. 1973. Subjective probability and its measurement. *Journal of the Royal Statistical Society: Series A (General)* 136, 1

(1973), 21–42.

[20] Merlin Heidemanns, Andrew Gelman, and G. Elliott Morris. 2020. An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election. *Harvard Data Science Review* 2, 4 (oct 27 2020). https://hdsr.mitpress.mit.edu/pub/nw1dzd02.

[21] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besancon. 2021. Can visualization alleviate dichotomous thinking? Effects of visual representations on the cliff effect. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2021), 3397–3409.

[22] J. G. Hollands and Brian P. Dyre. 2000. Bias in proportion judgments: the cyclical power model. *Psychological review* 107, 3 (2000), 500.

[23] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 903–913.

[24] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[25] Harald Ibrekk and M. Granger Morgan. 1987. Graphical Communication of Uncertain Quantities to Nontechnical People. *Risk Analysis* 7, 4 (1987), 519–529. https://doi.org/10.1111/j.1539-6924.1987.tb00488.x

[26] Daniel Kahneman and Amos Tversky. 1979. *Prospect Theory: An Analysis of Decision under Risk*. Vol. 47. [Wiley, Econometric Society], 263–291. http://www.jstor.org/stable/1914185

[27] Alex Kale, Matthew Kay, and Jessica Hullman. 2021. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

[28] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 892–902.

[29] Matthew Kay. 2021. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151

[30] Matthew Kay. 2019. How Much Value Should an Uncertainty Palette Suppress if an Uncertainty Palette Should Suppress Value?. In OSF Preprints. https://doi.org/10.31219/osf.io/6xcnw

[31] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5092–5103. https://doi.org/10.1145/2858036.2858558

[32] Drew A. Linzer. 2013. Dynamic Bayesian Forecasting of Presidential Elections in the States. *J. Amer. Statist. Assoc.* 108, 501 (2013), 124–134. https://doi.org/10.1080/01621459.2012.737735

[33] Le Liu, Lace Padilla, Sarah H Creem-Regehr, and Donald H House. 2018. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 882–891.

[34] Tingting Liu, Xiaotong Li, Chen Bao, Michael Correll, Changehe Tu, Oliver Deussen, and Yunhai Wang. 2021. Data-Driven Mark Orientation for Trend Estimation in Scatterplots. Article 473, 16 pages. https://doi.org/10.1145/3411764.3445751

[35] Min Lu, Joel Lanir, Chufeng Wang, Yucong Yao, Wen Zhang, Oliver Deussen, and Hui Huang. 2022. Modeling Just Noticeable Differences in Charts. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 718–726. https://doi.org/10.1109/TVCG.2021.3114874

[36] Mark J. Machina and David Schmeidler. 1992. A More Robust Definition of Subjective Probability. *Econometrica* 60, 4 (1992), 745–780. http://www.jstor.org/stable/2951565

[37] Luana Micallef, Gregorio Palmas, Antti Oulasvirta, and Tino Weinkauf. 2017. Towards Perceptual Optimization of the Visual Design of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1588–1599. https://doi.org/10.1109/TVCG.2017.2674978

[38] John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *The computer journal* 7, 4 (1965), 308–313.

[39] Jamie R. Nuñez, Christopher R. Anderton, and Ryan S. Renslow. 2018. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLOS ONE* 13, 7 (08 2018), 1–14. https://doi.org/10.1371/journal.pone.0199239

[40] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. Uncertain judgements: eliciting experts' probabilities. (2006).

[41] Lace MK Padilla, Sarah H Creem-Regehr, and William Thompson. 2020. The powerful influence of marks: Visual and knowledge-driven processing in hurricane track displays. *Journal of experimental psychology: applied* 26, 1 (2020), 1.

[42] Lace MK Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. 2021. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology* 11 (2021), 579267.

[43] Daniel Pineo and Colin Ware. 2012. Data Visualization Optimization via Computational Modeling of Perception. *IEEE Transactions on Visualization and Computer Graphics* 18, 2 (2012), 309–320. https://doi.org/10.1109/TVCG.2011.52

[44] Ghulam Jilani Quadri, Jennifer Adorno Nieves, Brenton M. Wiernik, and Paul Rosen. 2022. Automatic Scatterplot Design Optimization for Clustering Identification. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–16. https://doi.org/10.1109/TVCG.2022.3189883

[45] Howard Rachlin, Andres Raineri, and David Cross. 1991. Subjective probability and delay. *Journal of the experimental analysis of behavior* 55, 2 (1991), 233–244.

[46] Ian T. Ruginski, Alexander P. Boone, Lace M. Padilla, Le Liu, Nahal Heydari, Heidi S. Kramer, Mary Hegarty, William B. Thompson, Donald H. House, and Sarah H. Creem-Regehr. 2016. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation* 16, 2 (2016), 154–172. https://doi.org/10.1080/13875868.2015.1137577

[47] Houshmand Shirani-Mehr, David Rothschild, Sharad Goel, and Andrew Gelman. 2018. Disentangling Bias and Variance in Election Polls. *J. Amer. Statist. Assoc.* 113, 522 (2018), 607–614. https://doi.org/10.1080/01621459.2018.1448823

[48] Nate Silver. 2012. The weatherman is not a moron. *The New York Times* 7 (2012), 2012.

[49] Nate Silver et al. 2018. 2018 Midterm Election Forecast. https://projects.fivethirtyeight.com/2018-midterm-election-forecast/senate/

[50] Nate Silver et al. 2020. 2020 Election Forecast. https://projects.fivethirtyeight.com/2020-election-forecast/

[51] Nate Silver et al. 2022. 2022 Election Forecast. https://projects.fivethirtyeight.com/2022-election-forecast/senate/

[52] Nathaniel Smith and Stefan van der Walt. 2015. A Better Default Colormap for Matplotlib. https://www.youtube.com/watch?v=xAoljeRJ3lU

[53] Stan Development Team. 2020. RStan: the R interface to Stan. http://mc-stan.org/ R package version 2.21.2.

[54] Stanley S Stevens. 1957. On the psychophysical law. *Psychological Review* 64, 3 (1957), 153–181.

[55] Danielle Albers Szafir. 2018. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 392–401. https://doi.org/10.1109/TVCG.2017.2744359

[56] Carl-Axel S Staël von Holstein. 1970. Measurement of subjective probability. *Acta Psychologica* 34 (1970), 146–159.

[57] Yunhai Wang, Xin Chen, Tong Ge, Chen Bao, Michael Sedlmair, Chi-Wing Fu, Oliver Deussen, and Baoquan Chen. 2019. Optimizing Color Assignment for Perception of Class Separability in Multiclass Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 820–829. https://doi.org/10.1109/TVCG.2018.2864912

[58] Sean Jeremy Westwood, Solomon Messing, and Yphtach Lelkes. 2020. Projecting Confidence: How the Probabilistic Horse Race Confuses and Demobilizes the Public. *The Journal of Politics* 82, 4 (2020), 1530–1544. https://doi.org/10.1086/708682

[59] Marcel Wunderlich, Kathrin Ballweg, Georg Fuchs, and Tatiana von Landesberger. 2017. Visualization of delay uncertainty and its impact on train trip planning: A design study. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 317–328.

[60] Min-ge Xie and Kesar Singh. 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* 81, 1 (2013), 3–39.

[61] Hang Zhang and Laurence Maloney. 2012. Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. *Frontiers in Neuroscience* 6 (2012). https://doi.org/10.3389/fnins.2012.00001