

In Dice We Trust: Uncertainty Displays for Maintaining Trust in Election Forecasts Over Time

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

Chloe Mortenson
Northwestern University
Evanston, IL, USA
chloemortenson2026@u.northwestern.edu

Erik C. Nisbet
Northwestern University
Evanston, IL, USA
erik.nisbet@northwestern.edu

Nicholas Diakopoulos
Northwestern University
Evanston, IL, USA
nad@northwestern.edu

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

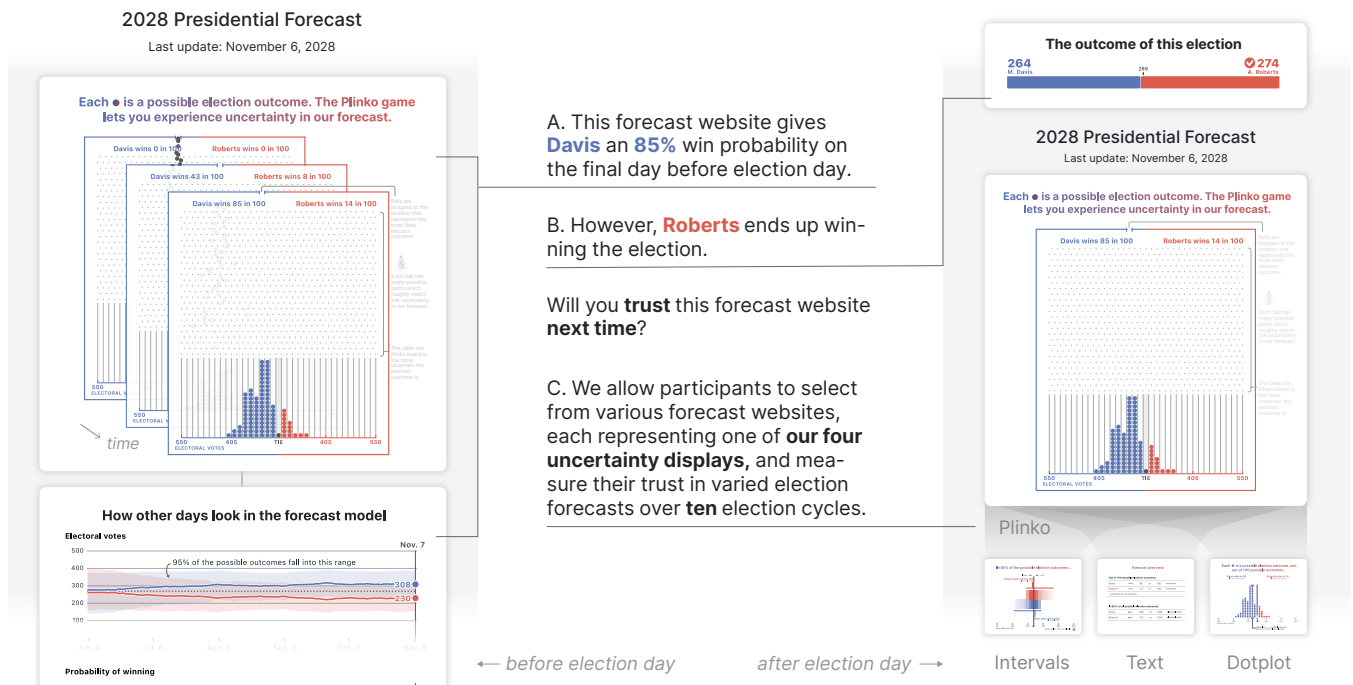


Figure 1: Illustration of an election cycle in our experiments: We show (A) a **simulated forecast** for the final day before election day and (B) subsequently **the election outcome**. In a professional-looking website interface, we showcase different (C) **uncertainty displays**: histogram intervals (intervals), a text summary (text), quantile dotplot (dotplot), and quantile plinko dotplot (plinko). Participants can choose a display/website, and their choices indicate part of their trust. We measure trust in simulated election forecasts throughout **ten** repeated election cycles. Note that the interface snapshots here are simplified. See Fig. 4 for the complete designs.

ABSTRACT

Trust in high-profile election forecasts influences the public's confidence in democratic processes and electoral integrity. Yet, maintaining trust after unexpected outcomes like the 2016 U.S. presidential election is a significant challenge. Our work confronts this challenge through three experiments that gauge trust in election forecasts. We generate simulated U.S. presidential election forecasts, vary win probabilities and outcomes, and present them to participants in a professional-looking website interface. In this website interface, we explore (1) four different uncertainty displays, (2) a technique

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0330-0/24/05

<https://doi.org/10.1145/3613904.3642371>

for subjective probability correction, and (3) visual calibration that depicts an outcome with its forecast distribution. Our quantitative results suggest that text summaries and quantile dotplots engender the highest trust over time, with observable partisan differences. The probability correction and calibration show small-to-null effects on average. Complemented by our qualitative results, we provide design recommendations for conveying U.S. presidential election forecasts and discuss long-term trust in uncertainty communication. We provide preregistration, code, data, model files, and videos at <https://doi.org/10.1145/3613904.3642371>.

CCS CONCEPTS

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

KEYWORDS

uncertainty visualization, trust, election forecasts, political communication

ACM Reference Format:

Fumeng Yang, Chloe Mortenson, Erik C. Nisbet, Nicholas Diakopoulos, and Matthew Kay. 2024. In Dice We Trust: Uncertainty Displays for Maintaining Trust in Election Forecasts Over Time. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 24 pages. <https://doi.org/10.1145/3613904.3642371>

1 INTRODUCTION

Probabilistic election forecasts provide an uncertain estimate of election outcomes over time. They typically draw upon opinion polling and fundamental data such as economic indicators. These forecasts serve various stakeholders' interests, and impact public sentiment [113], voter turnout [73, 108], as well as campaign strategies [66]. High-profile election forecasts, like FiveThirtyEight's [3] and The Economist's [5] U.S. presidential forecasts, come under intense scrutiny and can shape both domestic and international perceptions [28, 44]. Trust in these election forecasts is paramount, influencing the public's confidence in the democratic process and the integrity of the electoral system.

Yet, due to their inherent uncertainty, seeding trust in these forecasts is a significant challenge. Despite extensive literature on uncertainty communication and especially uncertainty visualization, the 2016 U.S. presidential election stands as a potent testament to this challenge. With a 71% win probability for Hillary Clinton on the final day [2], the election outcome for Donald Trump was unexpected for many, leading to public skepticism, distrust, and dissonance in both the forecasts and election outcomes [44, 103]—even though a 71% win probability indicates high uncertainty. Gaining public trust is not a one-time endeavor. It demands communication of uncertainty and calibration of outcomes over multiple election cycles. The U.S. electoral system, with its Electoral College, introduces even further complexities to communicating this uncertainty and building trust.

Our work confronts this challenge of maintaining trust in U.S. presidential forecasts. We conduct three experiments, each consisting of ten election cycles, and gauge how different presentations

of forecasts and outcomes affect people's trust over time. Under a cover story of a hypothetical 2028 U.S. presidential election, we first simulate U.S. presidential forecasts with varied win probabilities and outcomes—*correct* and *incorrect* (Sec. 3). We then construct an interface that resembles a professional forecast website (Sec. 4).

In this interface, we present four uncertainty displays (Fig. 1, each framed as a website), systematically adjust forecast distributions (*probability correction*, Fig. 3) [114], and depict outcomes in comparison to forecast distributions (*visual calibration*, Fig. 5). Using an incentivized voting task, we elicit participants' **attitudinal trust** (*perception of a forecast website*) and **behavioral trust** (*action of choosing a forecast website*) [16, 41] in these forecasts (Sec. 5), and analyze the data using Bayesian autocorrelation models (Sec. 6).

Specifically, we contribute:

- Design exploration for depicting U.S. presidential election forecasts and outcomes, including adopting and applying probability correction and visual calibration (Sec. 4);
- Quantitative results of people's trust in election forecasts over time, based on four uncertainty displays and simulated forecasts for U.S. presidential elections (Sec. 7);
- Qualitative results of why people trust a forecast website and how they utilize the uncertainty information presented on the website (Sec. 8);
- Design recommendations for conveying U.S. presidential election forecasts (Sec. 10).

Our sequentially preregistered online experiments recruited a total of 498 participants, with a balanced representation of gender and partisanship. The results show that a text summary (text, Fig. 1) and a quantile dotplot (dotplot, Fig. 1) gain the highest trust over time in repeated election cycles, substantially higher than trust in intervals (Fig. 1) and plinko (Fig. 1). Both probability correction and visual calibration show small-to-null effects on average, while partisan-motivated reasoning and forecast correctness have substantial effects on trust. These underscore the difficulty of fostering trust in probabilistic forecasts over time as a more general question. If you leave your umbrella at home when the weather forecast predicts a 29% chance of rain—and then get wet—will you still trust the forecast next time? Our research sheds light on the mechanics of trust maintenance in the face of probabilistic election forecasts.

1.1 Cover story

We aim to be able to provide design recommendations for conveying U.S. presidential election forecasts to the general public. Such forecasts are typically dynamic and updated throughout an election season, which often spans 155 days [3]. To capture this data generation process, we will simulate election outcomes and forecasts of those election outcomes across a similar timespan.

We set 2028 as the backdrop for our cover story—a presidential election year sufficiently distant from the time of this work to minimize contemporary biases, while being proximate enough to avoid a dramatic change in the U.S. political landscape. The two hypothetical presidential candidates are *M. Davis (Democrat)* and *A. Roberts (Republican)*, using common English last names and initials to mitigate race and gender implications [71, 93].

Participants navigate each election cycle characterized by two stages: **before** and **after** election day (Fig. 1). Given the electoral college’s function in U.S. elections, citizens technically vote for a slate of electors who pledge their electoral votes to that presidential candidate. Thus, we position participants within a designated U.S. state, which we will henceforth refer to as **the story state**.

Our primary focus is on the probability of winning the electoral college. Without loss of generality, we concentrate on *Davis’s* win probability ($p_{\text{DEM}} \approx 1 - p_{\text{REP}}$). We **vary** final-day probabilities of winning **the electoral college** and **fix** the win probability of **the story state**. Probabilities for other states are diverse but realistic. Participants experience forecasts of varied win probabilities and encounter different election outcomes. We present the final-day forecast for winning the electoral college as the website headline, followed by state details (Fig. 4). For more information about the U.S. political context, we direct readers to Sec. 2.2.

2 RELATED WORK AND BACKGROUND

Our work builds upon literature on uncertainty communication, trust, and voter turnout, including a recent study on election forecast visualizations [113]. To assist readers from different backgrounds, we also provide a brief primer to introduce the subjective probability correction [114] used in some of our experiments, as well as the basics of the U.S. electorate system and election forecasts.

2.1 Related work

Uncertainty displays. The existing literature developed textual, graphic, and audible [19] representations for uncertainty communication. Common displays include text summaries (e.g., 50 in 100) [81, 100], icon arrays [65, 111], intervals (e.g., error bars) [37, 99], ribbons [76], distributional plots (e.g., PDFs [72], CDFs [50], violin charts [65], histograms [114], and quantile dotplots [50, 61, 77, 82]), as well as animation (e.g., HOPs [70, 116] and Plinko [113]). They help in tasks like probability estimation [77, 111], transportation decision-making [50, 77, 82], and hurricane evacuation [87, 98]. In the realm of election forecasts, media outlets explored bee swarm plots [3], histograms [4], text summaries [5], and intervals [7]; and Yang et al.’s experiment during the 2022 U.S. midterm identified potential impacts of interval and quantile dotplots on people’s emotion and intention [113]. Uncertainty communication has been increasingly discussed in journalism [45, 68, 110], but remains difficult with audiences coming from diverse backgrounds [56].

Trust. Many research communities, such as AI/ML [22, 60], economics [26, 84, 96], political science [29, 95], and visualization [35, 47, 48, 53], have attended to trust through different lenses. As a fundamental aspect of social structures, trust determines how people interact with others and utilize new techniques. Various models and scales have been developed to operationalize and measure trust (e.g., [47, 89]) as well as longitudinal trust [43, 94]. We delve into two dimensions of trust—attitudinal and behavioral trust—to distinguish perception and action [16, 41]. We are inspired by the trust game [26, 96], an experimental economics exercise in which one player decides how much money to send to a second player (the first player’s trust), who receives a multiplied amount and then chooses how much to return (the second player’s trustworthiness).

Such behavioral outcomes reflect people’s genuine preferences, in contrast to sometimes misleading perceptual and proxy measures [31]. Yet, attitudinal trust can be a mediator for behavioral trust. In our experiments, each of these two measures collectively examines cognitive (e.g., rational evaluation) and affective trust (e.g., emotional bond) [48, 113] in data, forecast model, and visualization, as a lay audience typically does not distinguish between them.

Voter turnout. While not the main focus of the present work, our experiments involve a voting task. Voter turnout is an indicator of a functioning democracy, suggesting that citizens are engaged and believe in the system’s legitimacy. In political science and economics, the effects of election polls and forecasts on shaping voting behavior are long-standing topics [17, 21, 23, 36, 38–40, 49, 52, 73, 106, 108], producing mixed results. Often bandwagon effects [21, 38, 49, 105, 106] and pivotality (the importance of a vote) [27, 57, 108, 113] are of concern. Political science studies typically utilize national or local surveys [27, 49, 52, 106], while economics research often devises incentivized behavioral games [14, 23, 73, 108]. A behavioral approach is more suitable for our interests in behavioral trust. However, we limit our interpretation of the turnout results to the context of our specific experiments, avoiding unwarranted generalization.

2.2 Background

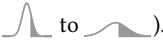
U.S. electoral system. At the time of this research, U.S. elections primarily revolve around a two-party system consisting of the left-leaning **Democratic (D)** and the right-leaning **Republican (R)** parties [91]. The most prominent national election is the quadrennial presidential election, which operates on an indirect voting system known as the **Electoral College** [80]. Each state is allocated a number of **electoral votes**, roughly proportional to its population, with a total number of 538 electoral votes. When casting votes for the president, citizens technically vote for a slate of electors who pledge their electoral votes to that presidential candidate. The winner in each state generally takes all of the state electoral votes, except for a few states using proportional systems [90, 91]. To win the presidency, a candidate must secure a majority of the electoral votes, 270 (269 is a tie, deferring the decisions to the Congress) [102]. The last day citizens can vote is **election day**, the first Tuesday of November. A **swing state** (“battleground state” or “purple state”) is where the two parties have similar voter support [24], making the election outcome uncertain. Because of the Electoral College, a small margin of victory in a swing state can give a candidate all of the state’s electoral votes.

U.S. election forecasts. Forecasts of presidential winners trace back to the 1930s [66, 86]. They typically cover the entire election season from June to November of a presidential election year, spanning about 155 days. Two major outlets publishing probabilistic forecasts for the U.S. elections are *FiveThirtyEight* (e.g., [2, 3, 8]) and *The Economist* (e.g., [5, 9]). Both present their forecasts in a fashion that starts with a summary of candidates’ forecasted electoral votes, followed by changes over time, state details, and relevant news articles (similar to our Fig. 4). They typically use uncertainty visualizations to convey the probabilistic forecasts and serve various viewers’ interests, ranging from satisfying curiosity to preparing related policies.

Uncertain displays of U.S. election forecasts. In addition to media news outlets’ practices, various scholarly works have explored representations for communicating election forecasts and polls (e.g., [67, 108]). Our work is inspired by the experiment conducted by Yang et al. during the 2022 U.S. midterm elections [113]. We adopt both their designs of annotations and choices of uncertain displays. However, we build upon their work and examine public trust over time in a decision-making context, highlighting the potential long-term impacts on democratic processes and electoral integrity. We also expand upon their choice of uncertainty displays to include a textual representation, and explore other uncertain communication techniques, as detailed below.

(Subjective) probability correction. Subjective probability describes people’s underlying belief in a probabilistic outcome [62, 88]. People may accurately report the exact win probability from a forecast website presented to them, but they tend to *internally shift* forecasted win probabilities towards 0 or 1 (e.g., misinterpreting a 71% chance of winning as a definite event). To account for these innate biases, subjective probability correction systematically adjusts the displayed probability distribution [114]. The core idea is to model subjective probability as a linear-in-probit (lpr) function of the true probability [117]:

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

and then invert this function to obtain the bias-corrected distribution (e.g., ). Intuitively, it undoes the bias occurring when going from true probability to subjective probability. The parameter α transforms the 50% focal point, and β scales the standard deviation of the distribution; the parameters are empirical point-estimation from a task eliciting subjective probability (e.g., betting on winners [20, 114]). This correction technique can substantially improve decision quality, compared to changing uncertainty representations [114]. In the present work, we refer to this technique as *probability correction* and apply it to U.S. presidential election forecasts in our current experiments.

3 SIMULATION

In order to conduct experiments with realistic election forecasting, we must first simulate U.S. presidential elections. We will use these simulations to generate both election outcomes and forecasts of those election outcomes. The probability of winning the electoral college (p_{DEM}) is computed from the predicted distribution of electoral votes, which is calculated based on the state vote share distributions (e.g., the state winner takes all of that state’s electoral votes). Therefore, the problem is equivalent to generating state vote share distributions for the Democratic candidate throughout the timeline.

3.1 Simulation goal

We will generate forecasts for **the entire election reason** (155 days) encompassing both the electoral college and all states. For the experimental proposes, we must have **varied** final-day win probabilities of **the electoral college**, i.e., p_{DEM} in {5%, 15%, ..., 95%}, and **fixed** win probability of **the story state**. We fix this state win probability to 50%, the most uncertain scenario, to engage participants in the experiments and motivate them to consult the

forecasts (also see Sec. 5.2 below). As such, the story state must be an authentic U.S. swing state, and we will also recruit participants from the swing states (see Sec. 5.6 below). The simulated forecasts should have properties similar to real-world presidential forecasts (e.g., variance, quality). Moreover, because we will also use these simulated forecasts to generate election outcomes, they need to capture the possibilities of both correct and incorrect election outcomes.

3.2 Simulation processes

We employ the data generation process proposed for forecasting the 2020 U.S. presidential election [64]. The core of this process is to assume vote share in different states fluctuates from day to day, constrained by a covariance matrix, which is calculated by Heidemanns et al. based on fundamental data (e.g., state population, demographics) [64]. Intuitively, if a Democratic candidate performs poorly in one blue state, the candidate is likely to perform poorly in other blue states. All state forecasts on the same day are sampled from the same multivariate normal distribution. Thus, we seek means and standard deviations of these multivariate normal distributions. Our simulation below is to (a) first generate a large set of forecasts using Monte Carlo simulation, guided by the properties derived from the past presidential forecasts, and (b) select those that meet the experimental requirements. We denote the state covariance matrix after the Cholesky decomposition by \mathcal{K} , a 51 by 51 matrix (50 states and Washington D.C.).

a. Generating forecasts for an election season. We first simulate elections and generate the **means** of state vote share distributions on each day (denoted by M_d) using Monte Carlo simulation:

$$M_d = M_0 + \sum_{i=1}^d \text{MVN}(0, \epsilon \cdot \mathcal{K}) \quad \text{for } d \in \{1..355\}$$

The initial means M_0 are states’ vote share percentages from the 2022 midterm elections (or 2020 elections for those who did not hold midterms in 2022). The day-to-day changes are drawn from a multivariate normal distribution (MVN) defined by a scaled covariance matrix ($\epsilon \cdot \mathcal{K}$). The constant ϵ is selected empirically to ensure the fluctuations in the resulting day-to-day win probabilities are less than 2%, aligning with FiveThirtyEight’s and The Economist’s 2020 forecasts. With a warm-up of 150 days to allow variability, we accumulate day-to-day changes and simulate $150 + 155 = 355$ days. We repeat this step 1,000,000 times, each producing all states’ vote share means over 155 days.

To generate forecasts, we must have the **standard deviations** of state vote share distributions on each day (Σ_d). The core idea is to scale the covariance matrix to obtain the desired margins (the width of 95% prediction intervals):

$$\begin{aligned} \Sigma_d &= \lambda_d \cdot \mathcal{K} \\ \lambda_d &= \alpha + \beta \cdot (155 - d)^2 \quad \text{for } d \in \{1..155\} \end{aligned}$$

The day-to-day scaling factor (λ_d) reflects the shrinking uncertainty approaching election day. This factor is defined by a two-parameter function varying with days [64]; the two parameters (α ; β) are selected via a grid search to constrain the margins of resulting

electoral votes. In alignment with FiveThirtyEight’s and The Economist’s 2020 forecasts, the first-day margin is between 230 and 240 electoral votes, while the last-day margin is between 152 and 156 electoral votes.

With means (M_d) and standard deviations (Σ_d), **state vote share distributions** on each day are defined by the following multivariate normal distribution:

All states’ vote share distributions on day $d \sim \text{MVN}(M_d, \Sigma_d)$

We draw 30,000 samples from each distribution. After this step, we have a set of reasonable forecasts, each containing predictions of all states (and thus electoral votes) over 155 days, and each day has 30,000 draws.

b. Selecting forecasts and outcomes for experimental purposes. First, we set the following criteria for win probabilities, which guarantee the availability of forecasts in the set generated above.

- The probability of winning the electoral college on the last day is in $\{5\%, 15\%, \dots, 95\%\} \pm 1\%$.
- The fluctuation of win probability in the last 7 days is less than 5%, which ensures stable trends and matches FiveThirtyEight’s and The Economist’s past U.S. presidential forecasts [2, 3, 5].
- One swing state has a win probability of $50\% \pm 3\%$, which will be the story state.

We draw samples from last-day forecast distributions to generate election outcomes, which leads to both *correct* (the outcome is the expected winner) and *incorrect* outcomes. We aim to control forecast quality and ensure a similar quality for all resulting forecasts, which will be used in the experiments. Given an outcome, we quantify forecast quality by the continuous ranked probability score (CRPS) [115], a commonly used scoring metric for probabilistic forecasts. It is defined as the distance between the CDF of forecast distribution and the step function of the outcome. Specifically, the 2020 presidential forecasts have a CRPS of 0.051 (FiveThirtyEight) or 0.053 (The Economist), meaning the outcome is about 45 electoral votes away from the forecast means. State forecasts usually have a CRPS of 0.01 to 0.07 [113], meaning the outcomes are 0.01 to 0.09 (vote share percentages) away from the forecast means. As such, we select the forecasts that contain last-day draws (i.e., election outcomes) to satisfy:

- The draw has a CRPS of 0.05 to 0.06 for electoral votes, and can be either correct or incorrect.
- The same draw has a CRPS between 0.01 and 0.02 for the story state, and can be either correct or incorrect.
- The other states in the same draw are reasonable: the number of correct states is 46, 47, or 48, as FiveThirtyEight and The Economist’s past forecasts have 46 to 49 correct states; the distance between the state outcomes and forecast means is always smaller than 0.075 (vote share percentages), and large states like California must be correct to ensure similar perceptions.

3.3 Simulation results

After the above steps, we have a set of U.S. presidential forecasts, each giving predictions of state vote share and electoral votes over

155 days. They have desired last-day win probabilities and election outcomes. Given any of these election outcomes, the corresponding forecast has a similar quality. Both the forecasts and their quality match past U.S. presidential forecasts. The story state in each forecast may vary, but it is one of possible U.S. swing states. We provide our code and the resulting data files in supplementary materials (▶ simulating elections).

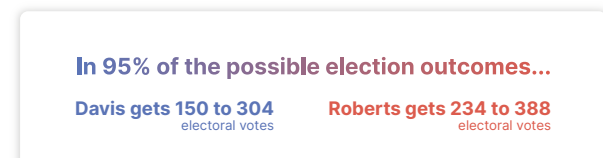
4 UNCERTAINTY DISPLAYS

For our experiments, we create a professional-looking forecast website and convey forecasts of both the electoral college and states. Following the 2020 presidential forecasts by FiveThirtyEight and The Economist, our website interface has a headline panel (Fig. 4A), a panel of changes over time (Fig. 4B), a state-level summary (Fig. 4C), and state-level details (Fig. 4D). Our experiments vary the headline displays (Figs. 1 and 2), use of probability correction (Fig. 3), and use of visual calibration (Fig. 5), all explorations of ways to maintain trust in election forecasts over time. The interface is color-blind-inclusive, and was tested in Chrome, Safari, and Firefox.

4.1 Headline uncertainty displays

The headline conveys the forecast of electoral votes. Our experimental choices are motivated by recent works on election forecasts [113, 114], particularly regarding the integration of annotations within visualizations. In a journalism context, appropriate annotations are essential; without them, it is nearly impossible for a lay audience to interpret a visualization appropriately. We adapt our approach from Yang et al.’s designs [113], aiming to choose each visualization along with its most suitable annotations. Despite inherent differences, we ensure consistency and comparability in annotations across various displays. Our annotations communicate the same probabilities and/or prediction intervals using consistent colors, fonts, and similar phrasing.¹ We also consider other works on uncertainty visualizations (e.g., [50, 72, 77]) and the current practice of media outlets (e.g., [3, 5, 7, 8]). We decide on four displays, representing the breadth of common design choices. We empirically pick 16 electoral votes as the bin size, which yields reasonable visualizations on a typical screen.

Histogram intervals (intervals, Fig. 2A) extend conventional representations [50, 72] and use illuminance to convey probability density inside intervals. Both The Economist’s 2021 German election forecasts [7] and Yang et al. [113] used a similar display; the latter work also showed that intervals have substantial effects on viewers’ emotion, trust, and intention. Following them, we annotate the 95% prediction intervals of electoral votes and their interpretation. For example,



¹Conducting a separate experiment on varied annotations could shed light on the effects of different annotations, which, however, falls beyond the scope of this work.

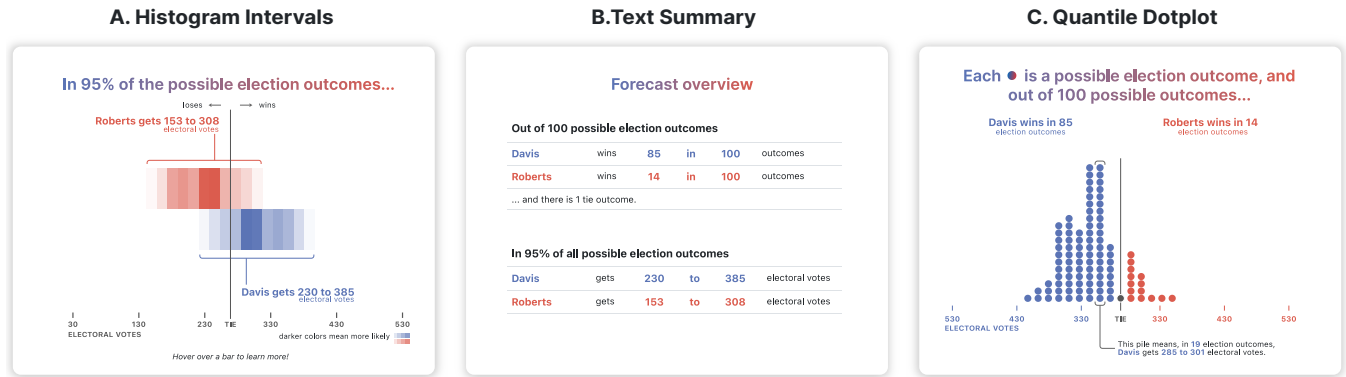
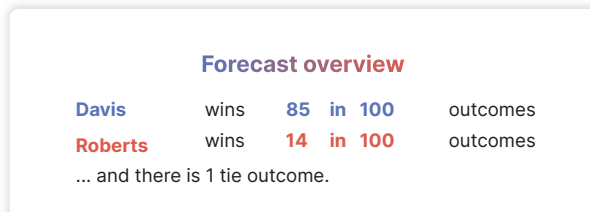
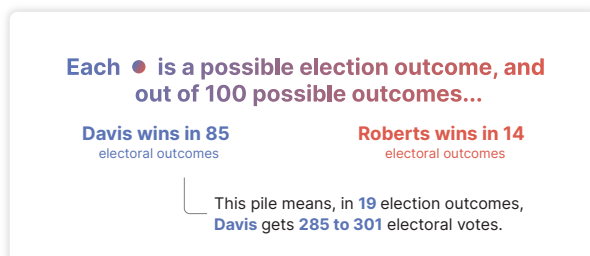


Figure 2: Examples of uncertainty displays: (A) histogram intervals (intervals), (B) a text summary (text), and (C) quantile dotplot (dotplot), all showing an 85% win probability for Davis. We present plinko and thumbnails in Fig. 1.

Text (text, Fig. 2B) summarizes both win probabilities and 95% prediction intervals. The Economist’s 2020 U.S. presidential forecasts [5] used the same representation. Other scholarly works compared a text summary with visual representations [50, 108], obtaining mixed results. Following the literature [100, 113], we convey win probabilities in a frequency style. For example,

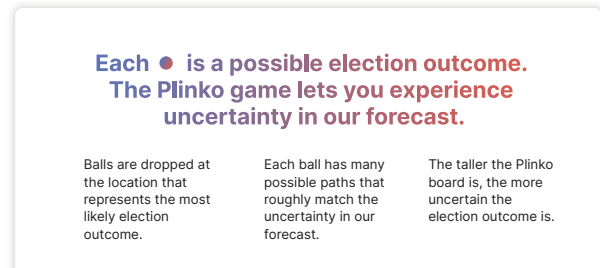


Quantile dotplot (dotplot, Fig. 2C) discretizes a probability density function [77]. FiveThirtyEight used a similar beeswarm plot for the 2020 U.S. general [3] and 2022 midterm [8] elections. Yang et al. explored both a single quantile dotplot and dual quantile dotplots [113], and found a single quantile dotplot had strong effects on emotions and trust in forecasts. Other works also support that quantile dotplots prompt understanding of uncertainty in decision tasks [50, 72, 82]. Following previous designs [113], our dotplot has 100 dots, and we annotate the meaning of a dot, the win probabilities, as well as the most likely outcome:



In consideration of a tie, we slightly modify the design and place tie outcomes in a separate bin (Fig. 2C).

Plinko quantile dotplot (plinko, Fig. 1) uses an animated physical analogy to depict the data generating process, formally introduced by Yang et al. [113]. The core concept is similar to the Galton Board [55], which approximates a forecast distribution using a Binomial distribution with a shifted mean. The Binomial distribution resembles a series of Bernoulli distributions, and a ball bounce on a peg represents each Bernoulli distribution. The variance of the Binomial distribution determines the height of the Plinko board. In Yang et al.’s results, Plinko quantile dotplots qualitatively improved viewers’ understanding of uncertainty, but slightly undermined trust in election forecasts [113]. The latency caused by animation appears detrimental to user experience and subsequently their trust [113]. As such, we make the following modifications: automatically starting the animation, shortening their duration from 60 seconds to 20 seconds, and animating only the election outcome in the post-election stage (Fig. 5C, also see Sec. 4.3). Our annotations simplify the cited work:



We provide videos in supplementary materials (📺 interface demo) to demonstrate the animation and annotations.

4.2 (Subjective) probability correction

One way to maintain trust in election forecasts might be to help people appropriately interpret probabilities (e.g., let them believe a 71% chance is a less likely event). As briefed in Sec. 2.2, probability correction is a technique that adjusts the displayed distribution to account for biases in people’s beliefs about the winner [114]. A probability correction relies on two parameters (α, β) that are estimated from empirical experiments. The original work provides

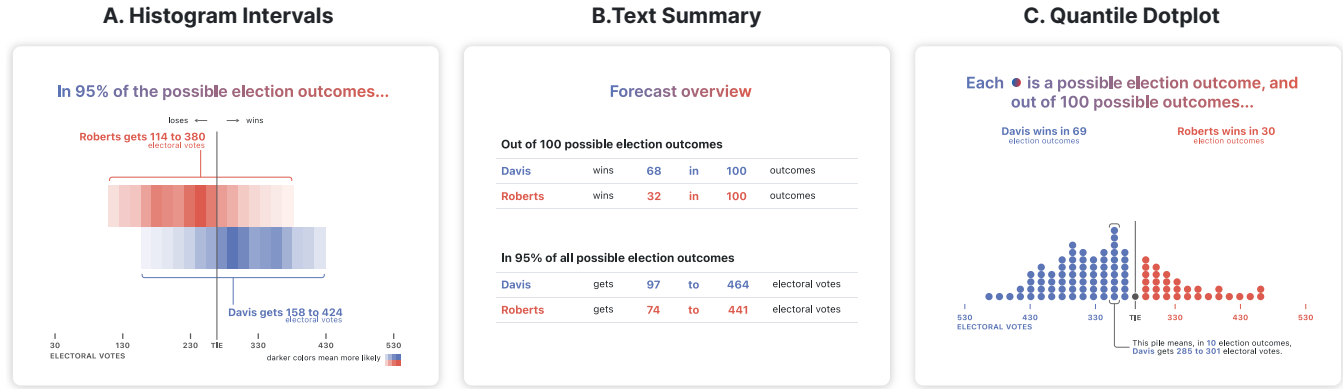


Figure 3: Examples of probability correction. The three examples here show our probability correction for an actual win probability of 85%, corresponding to Fig. 2. We eliminate plinko for the consideration of probability correction.

the parameters for histograms and a point-estimation text description [114]. We face several issues when applying this technique to U.S. presidential elections, elaborated below.

Correction parameters. The original work does not provide correction parameters for intervals, dotplot, and text (our text is an interval-style description). To obtain the parameters, we conducted new experiments. We used the same experimental materials and analysis code obtained from the original authors. We ran a simulation and found a sample size of 50 participants per condition yielded sufficient precision. We collected empirical data for the three displays, recruiting 53, 55, and 48 Prolific participants,² respectively. Since this probability correction technique amplifies variance, resulting in an excessively space-consuming board for plinko, we exclude plinko for consideration.

Ambiguous viewers. The original probability correction is tailored to scenarios where a viewer is primarily concerned with their preferred candidate’s win probability, a left- or right-tail probability of the forecast distribution. Such corrections are asymmetric, slightly different depending on whether the viewer prefers the Democratic or Republican candidate. However, in a journalistic context, we may not know the viewer’s preferred candidate, thus, we need a symmetric correction agnostic to an assumed preference. We modify the original model and fix the intercept parameter (α) to 0.³ This yields a symmetric probability correction, roughly equivalent to scaling the standard deviation of the forecast distribution by β . The final parameters (β s) for our probability correction are 1.55 (intervals), 2.02 (dotplot), and 2.14 (text), respectively.

Other considerations. Because our emphasis is on conveying the forecast of electoral votes, we apply probability correction only to this distribution (the headline), preserving uncorrected state-level forecasts. However, our scope encompasses the entire election season, spanning 155 days, requesting a transformation of the forecast of each day. Lastly, following suggestions from the original

work, we add a note into the interface to transparently apply a correction:

To adjust for how people typically discount uncertainty in forecasts, the standard deviation of electoral votes shown is approximately 2x that of our model. We hope that this will better explain the uncertainty for the reader.

We show examples of the probability-corrected distributions in Fig. 3. Note that we do not apply a correction to any election outcome. Readers can find our experimental materials and code in supplementary materials (📎 prob correction).

4.3 Visual calibration

Another way to maintain trust in election forecasts can be to release a post-election model calibration, a visual comparison between the actual outcome and the probabilistic forecasts, to enable viewers to appropriately judge the forecast quality. Yang et al. presented their calibration for state outcomes of the 2020 U.S. midterm elections [113]; meanwhile, FiveThirtyEight also publishes aggregated calibration on a separate page [10, 11]. With simulated forecasts, we can formally investigate the effects of visual calibration.

We adopt the visual calibration designed by Yang et al. [113] and represent an outcome as part of the forecast distribution (Fig. 5). These visual comparisons are annotated in a distinct green color. For intervals, we annotate the outcome onto the intervals using \hat{h} . For dotplot, we display the outcome as one of the dots using \circ , and annotate its meaning. For plinko, Yang et al. animated an election outcome along with other dots, replaying the entire animation; we modify their design and animate only the election outcome, shortening the duration to 5 seconds. For text, we reiterate the election outcomes to ensure consistency with the other displays in the experiments.

Additionally, we present visual calibration of state-level election outcomes. As the state outcomes are not of primary interest, we use the same design across all variants of headline displays (Figs. 4C and D). For state-level summary (snake charts; Fig. 4C), we keep the same layout, altering colors to represent the outcomes.

²Small variations in sample sizes do not substantially influence the resulting adjusted distributions. The actual sample sizes fluctuate with recruitment discrepancies on Prolific and imperfect assignments in Qualtrics.

³The parameters provided in the original work [114] are: $\alpha = -0.34$, $\beta = 2.44$ for text, and $\alpha = -0.34$, $\beta = 1.58$ for histograms. The α values were already close to 0.

I. Before election day

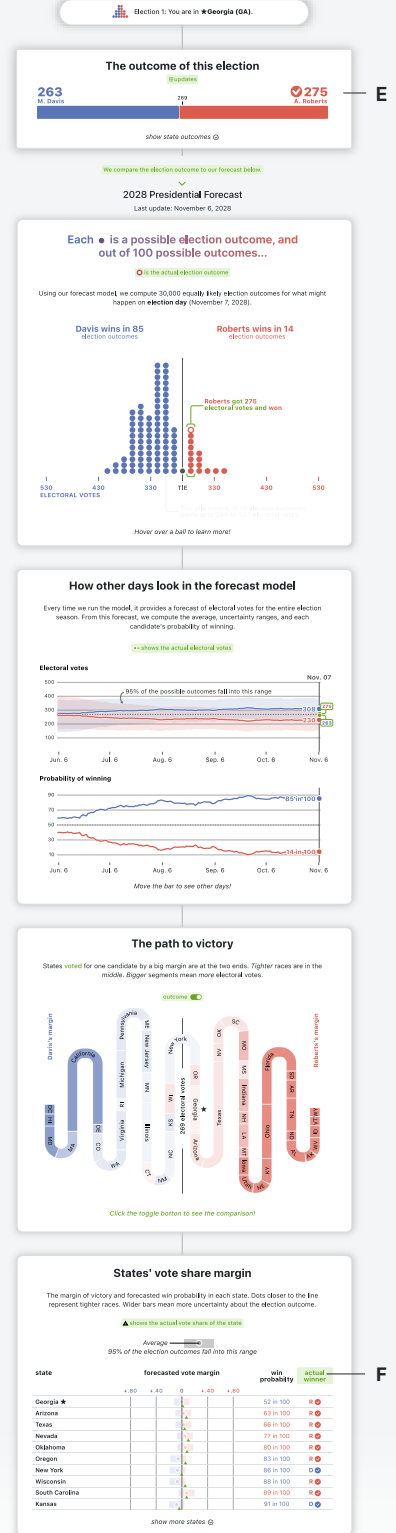
Snapshots of the interface

AN EXAMPLE OF NO CORRECTION, CALIBRATION

In each election cycle, we ask (I) a **voting** question before election day and (II) **trust** questions after election day, displayed in pop-out windows overlaid on the interface.



II. After election day calibration



This headline panel shows a display of forecasted electoral votes.

In all three experiments, participants choose the display for the next forecast via the options provided.

In Experiments 1 and 2, we vary **probability correction** and **calibration** for this panel between participants.

When not showing calibration, this panel stays the same from pre- to post-election (I to II).

This panel displays changes for electoral votes over 155 days. When applying **probability correction**, the ranges and probabilities displayed change accordingly.

When showing visual calibration (right), we annotate the election outcomes.

When not showing visual calibration, this panel stays the same when going from pre- to post-election (I to II).

This panel is a summary of state-level forecasts, inspired by FiveThirtyEight's design. Tighter races are in the middle.

The **visual calibration** (right) recolors all states by the election outcomes.

When not showing visual calibration, this panel stays the same when going from pre- to post-election (I to II).

This panel shows state-level details, the forecasted vote margins.

The visual calibration (right) annotates the state outcomes onto the intervals.

E The election outcome is always displayed (II) in a post-election stage regardless of calibration.

F When not showing calibration, this panel adds only **the actual winner** column when going from I to II.

The headline in the next election depends on participants' choice. →

Figure 4: Each forecast/election cycle has two stages: (I) Before election day, this example of dotplot shows an 85% probability of winning the electoral college (without probability correction). **(II) After election day**, participants are informed of (E) the election outcome; this example of visual calibration depicts the incorrect election outcome in relation to the forecast distribution.

Given the widening partisan divide among U.S. electorates [13, 29], partisan priming becomes necessary. We enforce a choice of the preferred candidate, including an option labeled “other candidate, not from the two major parties”. Their preferred candidate is who they would vote for (or not) throughout the experiment. This design ensures consistency and prevents insincere voting, as participants might switch allegiances merely to win a reward—a scenario unlikely in real elections [6]. This approach also simplifies experiments and enables us to collect more reliable turnout data, which is essential for addressing questions relevant to democratic participation [90]. We also inform participants of voting incentives. They start with a balance of 100\$, and this balance at the end of the experiment will be converted to a bonus at a 10:1 rate (e.g., 10\$ means 1 USD). The rules are

Voting has a cost -1\$; if your candidate wins the electoral college, you gain +5\$; if they lose, you lose -2\$.

Abstaining costs nothing 0\$; if your candidate wins the electoral college, you gain -2\$; if they lose, you lose +3\$.

These rules ensure (1) the same expected rewards regardless of whether participants consistently vote or abstain, preventing any unintended encouragement towards either action, and (2) a realistic turnout rate. By setting the voting cost to be a fifth of the winning reward, we aim for a turnout rate that aligns with the typical range observed in U.S. elections (40% to 70%) [14, 108]. We also pilot several rounds with different incentives and without incentives (asking participants to behave “naturally”). The results indicate that minor variations in incentives do not significantly impact turnout, but having no incentives leads to an unrealistic high voter turnout near 100%. We also anticipate various voting strategies [83], such as voting out of a sense of obligation.

5.3 Forecast finalization and assignment

From the forecasts generated in Sec. 3, we select ten with varied win probabilities and outcomes. We present the same ten forecasts to all participants and counterbalance them across participants to ensure experimental validity:




- Each participant is exposed to two incorrect forecasts, one for each party, which roughly matches the historical accuracy (~80%) of forecasting elections [78].
- The election outcomes presented to participants are consistent with forecasted win probabilities for both the electoral college and the story state. For instance, if the forecasted win probability is 50%, roughly half of the participants see an outcome of Davis winning.
- The order of the ten forecasts is randomized; however, the first forecast is always correct, matching the most recent U.S. presidential forecast in 2020 to provide participants with a sense of realism.
- The electoral college winner is balanced: each participant sees five wins each for Davis and Roberts.



- The story states are balanced: five total states, each occurring twice. These five states are Arizona, Georgia, Kansas⁴, Wisconsin, and Nevada.


Order randomization is generated onsite; the others are satisfied via a constraint programming algorithm. This algorithm outputs ten different specifications, each containing the ten forecasts and randomly assigned to participants. The exact win probabilities of the ten forecasts are {5%, 15%, 24%, 35%, 44%, 56%, 65%, 75%, 85%, 94%}.

5.4 Experimental design

We adopt a sequential design process, illustrated in Fig. 6. Given the unpredictability of the results, this approach allows for iterative refinement, wherein each experiment builds upon the findings of the preceding one, culminating in our final recommendations. We initiate with Experiment 1 and use the preceding results to inform the design and preregistration of subsequent experiments. Participants from the pilot studies, earlier experiments, or any related studies we conducted are excluded from subsequent experiments. We provide our preregistration in supplementary materials (📄 preregistration).

Experiment 1. We start with a larger experiment varying both probability correction and calibration. To eliminate ambiguity, we let participants choose from only three displays: intervals , text , and dotplot . In other words, both probability correction and calibration are between-subjects variables, and each participant is assigned to one of the four combinations {no correction, correction} × {no calibration, calibration}. We exclude plinko because it is difficult to apply probability correction to (see Sec. 4.2 above), and its inclusion would complicate result interpretation. We use pilot data to generate synthetic data for estimating model precision, and decide on 300 participants.

Experiment 2. We then compare the “winners” of Experiment 1 with plinko. The results of Experiment 1 suggest that, on average, participants choose text slightly more than dotplot, both ahead of intervals. We had piloted with text, dotplot, and plinko. Nevertheless, the results raise a concern about the independence of irrelevant alternatives [25]. That is, the presence of a third option may unequally decrease the probability of selecting the other two. Thus, we reduce the options to dotplot  and plinko  for a precise comparison between the two visual displays, varying no calibration and calibration between participants. The number of conditions is one-third of that in Experiment 1, and therefore we decide on 120 participants and obtain 119, as Prolific sometimes has discrepancies in participant numbers.

Experiment 3. Combining the results of Experiments 1 and 2, dotplot gains the highest trust among the three visual displays. However, on average, the effects of correction and visual calibration are small and may be subject to individuals’ characteristics (e.g., partisanship). Therefore, Experiment 3 disambiguates text  and

⁴While Kansas might not be considered a conventional swing state, in both 2018 and 2022 gubernatorial elections, the popular votes are very close (e.g., 49.54% vs. 47.33%). The state is likely to be a swing state in the future.

The design flow of our sequentially preregistered experiments

Experiment 1

Exploring a larger set

300 participants

VISUALIZATION OPTIONS



intervals



text



dotplot

BETWEEN-SUBJECTS VARIABLES

correction calibration

Key results

1% to 99% CIs

Average probabilities of selecting a visualization

TAKEAWAYS



Average ratings over all forecasts



- Trust in **dotplot** and **text** seems similar, higher than trust in **intervals**.
- The presence of a third option may unequally affect the probability of selecting the other options (IIA).

Experiment 2

A head-to-head comparison

119 participants

VISUALIZATION OPTIONS



dotplot



plinko

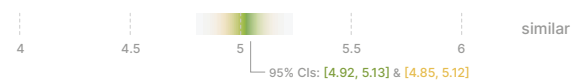
BETWEEN-SUBJECTS VARIABLE

calibration

Average probabilities of selecting a visualization



Average ratings over all forecasts



- Participants choose **dotplot** over **plinko**, but attitudinal trust is similar between the two.
- Calibration may have small-to-null effects.

Experiment 3

The final run-off

79 participants

VISUALIZATION OPTIONS



text



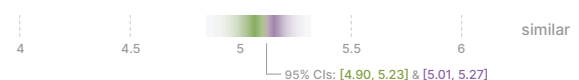
dotplot

Both are with calibration.

Average probabilities of selecting a visualization




Average ratings over all forecasts



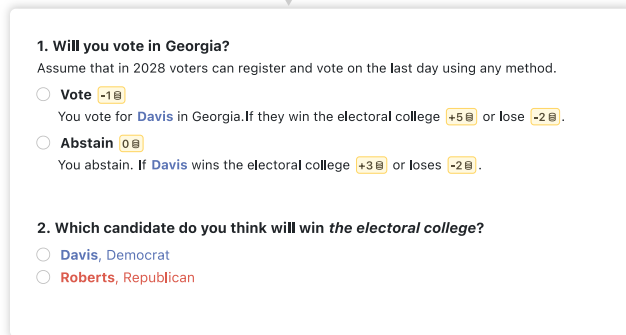
- Participants trust **text** more than **dotplot**, which is the most trusted visual display in the experiments.

Figure 6: The design flow and key results of our sequentially preregistered experiments, each building upon the preceding results. Overlapped intervals are blended using ggblend [74].

dotplot  in the presence of visual calibration. The number of conditions is further halved. We decide on 81 participants and obtain 79, excluding two participants who were retaking the experiment.

5.5 Procedure and interface

After consent, demographics, cover story, and selecting their candidate (e.g., *Davis*, see Sec. 5.2 above), participants have a test trial to learn the association between icons and forecast websites. Then, in the first forecast, they are randomly assigned to one of the displays depending on the experiment. They complete ten forecasts, each having two stages. **Before election day**, they answer two questions.



1. Will you vote in Georgia?
Assume that in 2028 voters can register and vote on the last day using any method.

Vote -1
You vote for *Davis* in Georgia. If they win the electoral college +5 or lose -2.

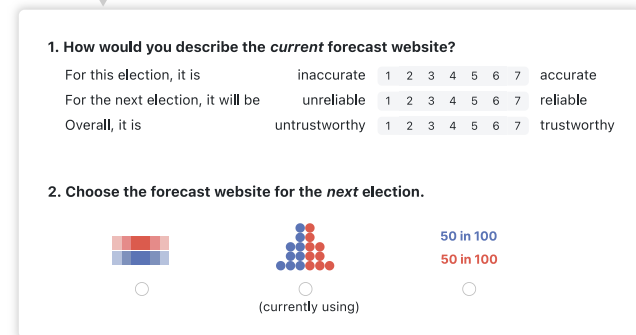
Abstain 0
You abstain. If *Davis* wins the electoral college +3 or loses -2.

2. Which candidate do you think will win the electoral college?

Davis, Democrat

Roberts, Republican

After election day, they report attitudinal and behavioral trust.





1. How would you describe the *current* forecast website?

For this election, it is inaccurate 1 2 3 4 5 6 7 accurate

For the next election, it will be unreliable 1 2 3 4 5 6 7 reliable

Overall, it is untrustworthy 1 2 3 4 5 6 7 trustworthy

2. Choose the forecast website for the *next* election.

  50 in 100
50 in 100

 (currently using)

The “which candidate...” question is a sanity check. Participants must see the headline panel before answering the questions in a pop-out window, and they can hide the window to refer back to the forecast at any time. The wording in the final forecast is slightly modified to fit the scenario. After ten forecasts, they respond to four open-ended questions, such as how they vote and choose a forecast website.

5.6 Participants and recruitment

We recruit all participants from Prolific.com and screen them based on their profiles reported to Prolific. The three experiments use the same screeners. We request participants who live or lived in the U.S. swing states: Oregon, Arizona, Nevada, Minnesota, Colorado, Ohio, Michigan, Wisconsin, Maine, New Hampshire, Pennsylvania, North Carolina, Georgia, Florida, and Texas. Because the turnout for the 2020 U.S. presidential election is approximately 66% [12], we balance partisanship and recruit participants who voted for *Joe Biden*, *Donald Trump*, and who did not vote at a rate of 1:1:1,

balancing gender for each. The exception is that Experiment 3 requests 28, 28, and 26 participants to have gender balance for an odd number of total participants.

We provide demographic breakdowns in supplementary materials (📎 data and analyses). Each participant is compensated 4 USD for their time. The mean completion time is 19.21, 22.45, and 21.28 minutes for the three experiments; and the mean bonus is 1.45, 1.52, and 1.51 USD. The accuracy of the winner question is 81.67%, 84.45%, and 86.08%. The study was approved by the Institutional Review Board (IRB) at Northwestern University (#STU00216162-MOD0002).

6 QUANTITATIVE ANALYSES

Because each participant undertakes ten forecasts, their responses likely depend on each preceding forecast. Therefore, we model their trust responses using Bayesian autocorrelation models. We use similar model specifications for all experiments and begin with describing Experiment 1 below, followed by the modifications for Experiments 2 and 3 to account for the different experimental designs. All model specifications and priors are preregistered.⁵

We used Rstan [104], CmdStanR [54], posterior [32, 107], and tidybayes [75] for our implementation. We inspected sampler transitions (treedepth, divergences, and E-BFMI), effective sample size, and R-hat values, all of which were satisfactory. The data, R and stan code, and model files are available in supplementary materials (📎 data and analyses).

6.1 Behavioral trust

Behavioral trust models the observations of participants’ choices of headline displays/visualizations. The core idea is to maintain a latent trust variable for each participant’s trust in each visualization over time, including an initial state indicating prior beliefs before the first forecast. Therefore, this latent variable has 11 dimensions (i.e., forecast t corresponds to time $t + 1$). After each choice, this latent variable is updated in accordance with the visualization of the current forecast and experimental variables. This latent variable generates each choice through a Multinomial distribution, except for the first forecast, where visualizations are assigned randomly. Corresponding to this latent variable, behavioral trust is measured by participants’ probabilities of selecting a particular visualization.

More formally, we express our model specification as follows.

⁵We preregistered more than one specification and a protocol to explore them. Because we find the primary specification satisfying, we present it here and use it for result reporting.

Observation distribution

The visualization VIS seen by PARTICIPANT i at time $t + 1$ (forecast t) is given by a **Multinomial** distribution, defined by the probabilities of seeing each visualization $\text{Pr}(\text{VIS}_{i,t+1})$. These probabilities are transformed from the latent trust \mathcal{H} of each participant i , visualization v , and time $t + 1$ via a **SOFTMAX** function.

$$\text{VIS}_{i,t+1} \sim \text{Multinomial}(\text{Pr}(\text{VIS}_{i,t+1} \mid \mathcal{H}_{i,t+1,\cdot}))$$

$$\text{Pr}(\text{VIS}_{i,t+1} \mid \mathcal{H}_{i,t+1,\cdot}) = \text{softmax}(\mathcal{H}_{i,t+1,\cdot})$$

Initialization

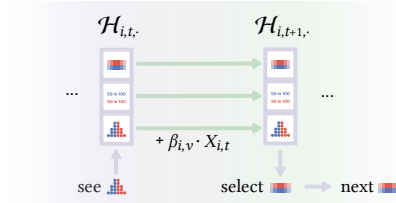
The latent trust before the first forecast is initialized to \mathcal{H}^* . The first dimension is set to 0 for the identifiability of the softmax function.

$$\mathcal{H}_{i,1,\cdot} = [0, \mathcal{H}_{i,2}^*, \mathcal{H}_{i,3}^*]$$

Updating

For participant i and forecast t , their latent trust $\mathcal{H}_{i,t+1,\cdot}$, which decides their choice for the next forecast, is updated for the current visualization $\text{VIS}[i, t]$ based on the current predictors $X_{i,t}$ and their coefficients $\beta_{i,\text{VIS}[i,t]}$. The other entries of latent trust remain the same. In Pinheiro-Bates notation [101], we define X as correctness:calibration + correctness:prob_correction + partisanship, indicating the effects from forecast correctness, visual calibration, probability correction, and partisan leaning (contrast: -5, 0, 5), as well as the interaction between them.

$$\mathcal{H}_{i,t+1,v} = \begin{cases} \mathcal{H}_{i,t,v} + \beta_{i,v} \cdot X_{i,t} & \text{if } v = \text{VIS}[i, t] \\ \mathcal{H}_{i,t,v} & \text{otherwise} \end{cases}$$



Hierarchical priors

We anticipate each participant i to have their own initial trust and coefficients for updating for each visualization v .

Indices for Experiments 1, 2, and 3

$N_{\text{PARTICIPANT}}$ is 300, 119, and 79, respectively, and N_{VIS} is 3, 2, and 2, respectively. N_{FORECAST} is always 10.

Modifications for Experiments 2 and 3

We modify the predictors (X) used in updating. Experiment 2 varies only visual calibration, and has correctness:calibration + partisanship, and Experiment 3 always shows calibration, and therefore has correctness + partisanship. Also, both experiments provide participants with two visualization options; the Multinomial distribution reduces to a Bernoulli distribution. However, we use the same expression and code for consistency.

$$\mathcal{H}_{i,v}^* \sim \text{Normal}(\mu_{i,v}^*, \sigma_{i,v}^*)$$

$$\mu_{i,v}^* \sim \text{Normal}(0, 3)$$

$$\sigma_{i,v}^* \sim \text{Exponential}(1)$$

$$\beta_{i,v} \sim \text{Normal}(\mu_{i,v}, \sigma_{i,v})$$

$$\mu_{i,v} \sim \text{Normal}(0, 1)$$

$$\sigma_{i,v} \sim \text{Exponential}(1)$$

$$i \in \{1..N_{\text{PARTICIPANT}}\}$$

$$t \in \{1..N_{\text{FORECAST}}\}$$

$$v \in \{1..N_{\text{VIS}}\}$$

6.2 Attitudinal trust

Attitudinal trust models the observations of the Likert responses. Following the analysis guideline for Likert format data, we take the mean of the three Likert questions as the observations of attitudinal trust [30, 63]. The model is substantially similar to that of behavioral trust above, except that a Normal distribution with a latent mean and standard deviation generates rating observations, defined below.

Observation distribution

We expect Rating by participant i in forecast t comes from a **Normal** distribution. Its mean is given by the latent attitudinal trust $\mathcal{P}_{i,t+1,v}$ at that time $t + 1$ for the participant and visualization $\text{VIS}[i, t]$.

$$\text{Rating}_{i,t} \sim \text{Normal}(\mathcal{P}_{i,t+1,\text{VIS}[i,t]}, \tau_{i,t+1,\text{VIS}[i,t]})$$

Initialization

The latent attitudinal trust is initialized for each participant i and each visualization v .

$$\mathcal{P}_{i,1,v} = \mathcal{P}_{i,v}^*$$

Updates, hierarchical priors, and indices

The updates, priors, and indices are the same as the behavioral trust model, except for including a prior τ on the standard deviation of the Normal observation distribution. The predictors and modifications for Experiments 2 and 3 are the same as above.

$$\tau_{i,t,v} \sim \text{Exponential}(1)$$

$$\mathcal{P}_{i,v}^* \sim \text{Normal}(\mu_v^*, \sigma_v^*)$$

...

6.3 Voting decision

We use a similar model structure for voting observations but without temporal autocorrelation. Given the experimental designs, participants are most likely to vote based on win probabilities and their voting habits. We also modify the predictors because participants vote in the pre-election stage. Voter turnout is then measured by the probability that an average participant votes.

Observation distribution

We view each binary voting decision j arising from a Bernoulli distribution [23, 85], with the voting probability θ_j as a function of experimental variables as the predictors.

$$\text{Voted}_j \sim \text{Bernoulli}(\theta_j)$$

Hierarchical priors and indices

We use a similar hierarchical model structure, where each PARTICIPANT[j] has their own coefficient $\gamma_{i,v}$ for each visualization VIS[j]. However, the predictors are different, denoted by Z and defined as $\text{logit}(p_{\text{dem}}): \text{party_candidate} + \text{prob_correction}$ in the logit space. This means the actual win probability interacts with which candidate they vote for (the same candidate they select at the beginning of the experiment), and probability correction also affects voting decisions. We remove the prob_correction term for Experiments 2 and 3.

$$\text{logit}(\theta_j) = \gamma_{\text{PARTICIPANT}[j], \text{VIS}[j]} \cdot Z_{\text{PARTICIPANT}[j], \text{VIS}[j]}$$

$$\gamma_{i,v} \sim \text{Normal}(\mu_v, \sigma_v)$$

$$\mu_v \sim \text{Normal}(0, 1)$$

$$\sigma_v \sim \text{Exponential}(1)$$

$$j \in \{1..N_{\text{OBSERVATION}}\}$$

$$v \in \{1..N_{\text{VIS}}\}$$

7 QUANTITATIVE RESULTS

With the models and measures, we first report the results from the preregistered analyses. At the end of this section, we report the results from a non-preregistered exploratory analysis. Because we balance gender, partisanship, and experimental conditions, we base our interpretation on participants' averages and report the medians and 95% credible intervals (CIs; Bayesian analog to confidence intervals) in the format of median [lower, upper]. For those model coefficients, we transform them to the original scale to facilitate interpretation.

7.1 Behavioral trust (probability)

Participants' average trust (Fig. 7A). We observe a clear difference in participants' behavioral trust. Their initial trust in each visualization can be similar (Experiments 2 and 3). However, participants' choices of visualization tend to be stable after two forecasts and gradually converge (Experiments 1 and 3) or diverge (Experiment 2).

More specifically, in Experiment 1, people choose **text** and **dotplot** over **intervals** (in the last forecast: 0.44 [0.43, 0.46], 0.37 [0.36, 0.39], 0.18 [0.17, 0.20]). In Experiment 2, people choose **dotplot** over **plinko** (in the last forecast: 0.66 [0.63, 0.68], 0.34 [0.32, 0.37]). In Experiment 3, people choose **text** slightly over **dotplot** (in the last forecast: 0.56 [0.54, 0.59], 0.44 [0.41, 0.46]). On average, participants tend to choose and therefore trust **text** the most, slightly more than **dotplot**, and both more than **intervals** or **plinko**.

Model coefficients (Fig. 7B). Across the three experiments and visualizations, correctness and partisanship consistently increase participants' behavioral trust. That is, when a forecast is correct (cf. incorrect) or predicts their candidate winning (cf. the opponent candidate), it can substantially increase the probability of choosing a visualization (e.g., Experiment 1, **dotplot** 0.14 [0.03, 0.24]). However, the effects of partisanship are less consistent and sometimes diminish (e.g., Experiment 2 **dotplot** and Experiment 3

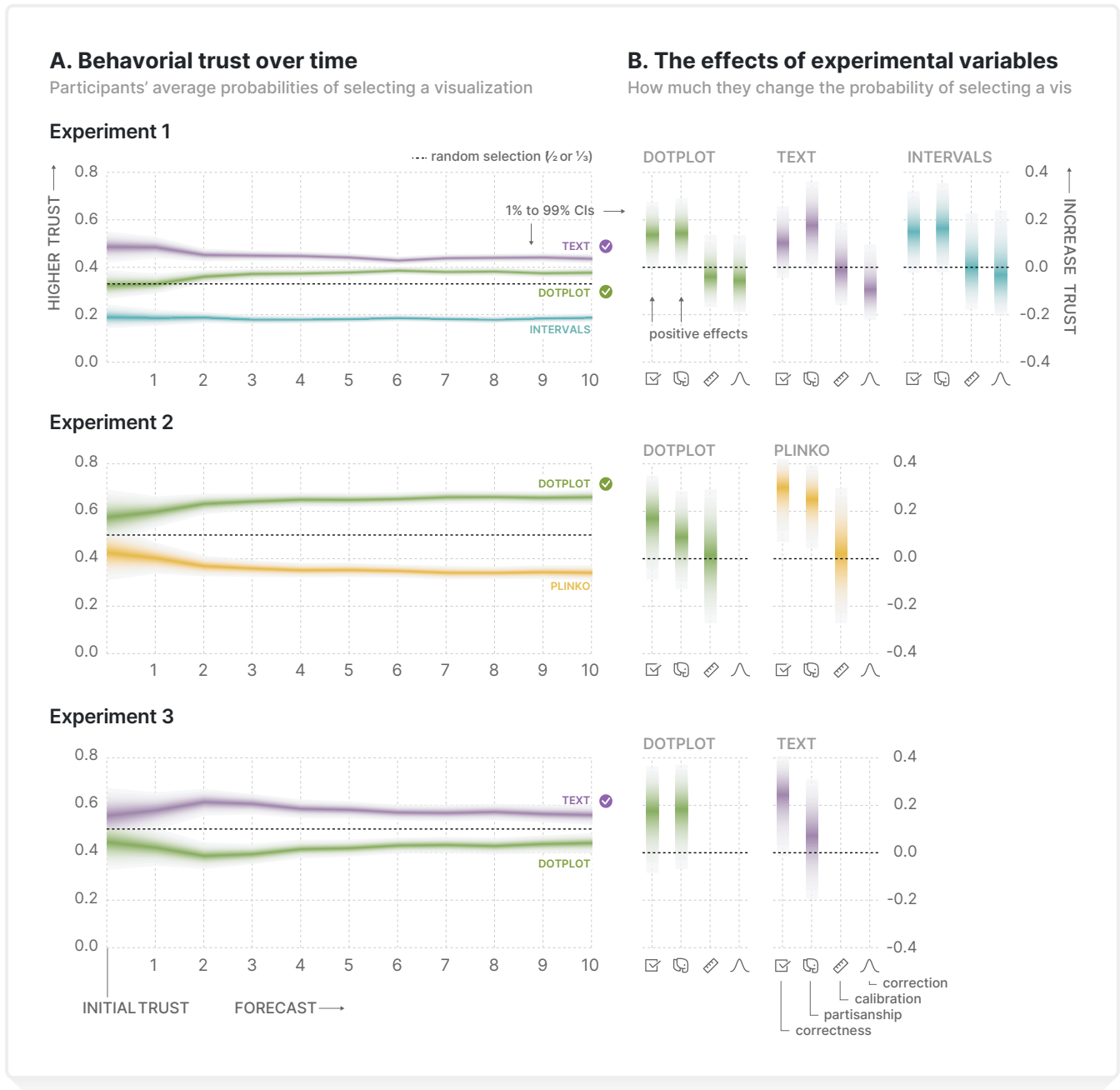


Figure 7: Behavioral trust over time. We show participants' average probabilities of selecting a visualization, including their initial trust before seeing any forecast, and how the experimental variables affect participants updating their trust over time.

text). On average, probability correction and calibration have non-conclusive effects on behavioral trust; we explore further in Sec. 7.4 below.

7.2 Attitudinal trust (Likert scale)

Participants' average trust (Fig. 8A). The results suggest that participants' perceptions of visualizations are similar, and the

effect sizes are generally small. Experiments 1 and 3 show their attitudinal trust in dotplot is similar to that in text, though text gains slightly higher trust in Experiment 3. However, in Experiment 1, their attitudinal trust in dotplot slightly improves over time, which contrasts with Experiment 3. In Experiment 2, attitudinal trust in dotplot and plinko are similar. Overall, the results of attitudinal trust corroborate with those of behavioral trust above—trust in

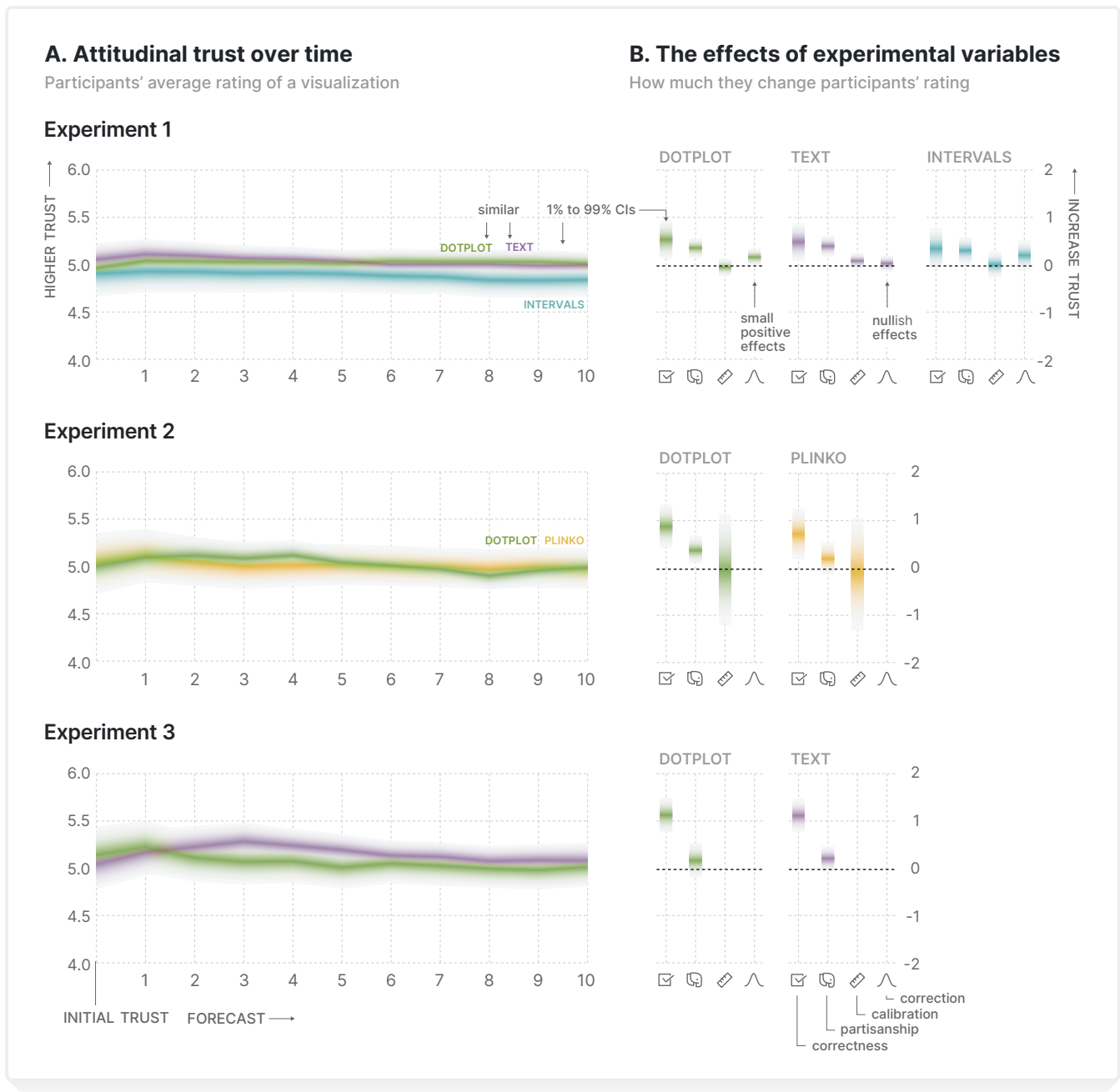


Figure 8: Attitudinal trust over time. We show the average participants' underlying rating of visualization, including their initial trust before seeing any forecast, and how the experimental variables affect the rating.

text and dotplot are similar, and higher than trust in intervals and potentially plinko.

Model coefficients (Fig. 8B). Across the experiments and visualizations, correctness consistently increases participants' attitudinal trust (e.g., Experiment 1, intervals 0.34 [0.014, 0.67]); the

partisanship effect is weaker and sometimes diminishes (e.g., Experiment 2, plinko 0.20 [-0.052, 0.46]; Experiment 3, dotplot 0.16 [-0.11, 0.44]). On average, in Experiment 1, correction has small positive effects on improving trust perceptions, especially for dotplot (0.16 [0.017, 0.30]).

7.3 Voter turnout (probability)

We also report voter turnout in Fig. 9, showing the probabilities of an average participant voting in the experiments. Our model contrasts the two parties based on the candidate participants select, and those selecting the “other candidate” are coded as 0, resulting in a uniform turnout distribution. Hence, we present the results of participants who support **Davis**, mirroring the results of participants who support **Roberts**.

Overall, we see that turnout increases as the win probabilities of their chosen candidate increase. If we take **text** as a baseline, then both Experiments 1 and 3 show that **dotplot** increases turnout over **text** (e.g., in Experiment 1, when $p_{DEM} = .45$, **dotplot** 0.65 [0.58, 0.72] and **text** 0.50 [0.41, 0.59]). However, both **intervals** and **plinko** can further increase voter turnout when the win probability is low (e.g., in Experiment 1, when $p_{DEM} = .15$, **intervals** 0.16 [0.11, 0.23], **dotplot** 0.087 [0.040, 0.17]), but may slightly decrease voter turnout when their win probability is high (e.g., in Experiment 1 when $p_{DEM} = .85$, **intervals** 0.89 [0.83, 0.94], **dotplot** 0.97 [0.94, 0.98]). Additionally, in Experiment 1, correction has small, non-conclusive effects on increasing turnout (Fig. 9A).

7.4 Non-registered exploratory analysis

Given the non-conclusive effects of probability correction and calibration on behavioral trust, we conduct a small-scale exploratory analysis based on participants’ choices in the last forecast/election cycle. Prior work identified gender [53] and partisan [113] differences in trust. In the present work, we find that the partisan divide is more consistent than any gender difference. As such, we first split participants based on the candidate they select in the experiments and report participants’ averages in Fig. 10.

We see that the choices of participants supporting the Democratic candidate appear different from those supporting the Republican and other candidates. In Experiment 1, Democratic participants choose **dotplot** slightly over **text** and **intervals** (Fig. 10A). In both Experiments 1 and 2, their choices of visualizations are getting closer in the presence of correction and/or calibration (Figs. 10A and B). We also observe a similar effect in Experiment 1 for Republican participants, but this effect diminishes in Experiments 2 and 3. These suggest that visual displays, correction, and calibration are preferred by Democratic participants.

Second, we further break down participants by their education levels, reported in Fig. 10D. We see a correlation between higher education levels and preferring **dotplot** in several subgroups, such as those Democratic participants in Experiment 1 and non-Democratic participants in Experiment 2, but this correlation seems not to hold universally.

8 QUALITATIVE RESULTS

To examine construct validity and gain further insights, we conduct a qualitative analysis of participants’ responses to two open-ended questions: how they decide to vote and choose a website. We did not distinguish between different experiments, and one coder analyzed all $498 \times 2 = 996$ responses. The coder started with open coding and then grouped the codes into axes [79]; the code assignments were not mutually exclusive. This qualitative analysis is also under our

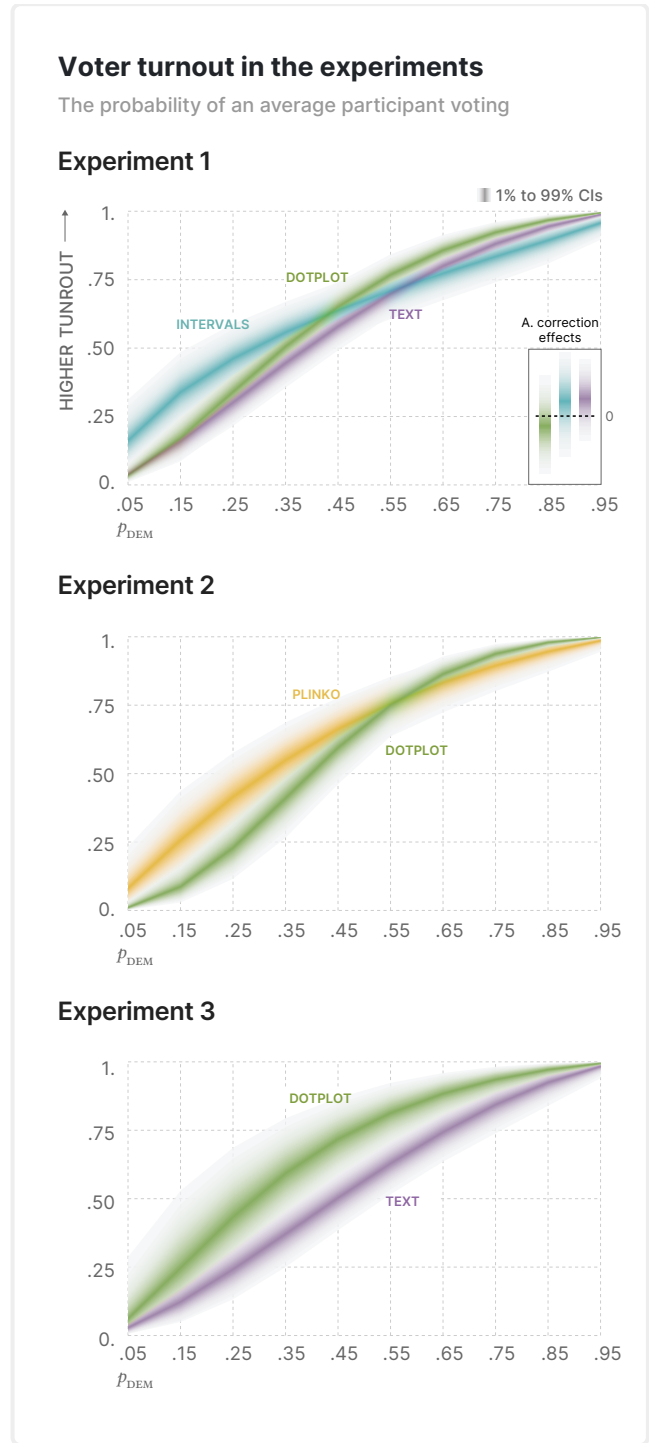


Figure 9: The turnout rate of those participants supporting **Davis** over their win probabilities p_{DEM} , mirroring the results of participants supporting **Roberts** and p_{REP} .

preregistration, and the codebook and assignments are provided in supplementary materials (📎 qualitative analysis).

8.1 Voting strategies (Tab. 1)

We find that 78% of participants use the information provided to make their voting decisions in the experiments. As expected, a portion of the participants always vote (13%), considered as “regular voters” in the literature [51]; and 3% of the participants (usually who support the “other candidate”) always abstain. In general, participants vote based on win probabilities (34%), forecasted winners (15%), or margin of victory (8%); some further consider state details (5%) and trends (1%). All suggest that the experiments prompt political thinking and interpretation of the forecasts, showing good construct validity. These also suggest what information was used by participants, providing clues for designing a forecast website.

8.2 Choosing/trusting a website (Tab. 2)

Participants may choose a website to finish the experiments efficiently (4%) or based on personal interests (ranging from 1% to 15%). Mostly, accuracy (22%), understandability (19%), readability (11%), preference (15%), and similar aspects capture the majority of responses. They can be interpreted as engagement in the experiments and a reflection of participants’ needs for election forecasts. Design factors such as aesthetics (3%) also affect some participants’ choices. These results suggest that participant behavioral trust is a result of both cognitive (e.g., forecast accuracy and visualization clarity) and affective trust (e.g., aesthetics and benevolence) [48]. The results also verify the construct validity of this work, and shed light on designing such an interface to convey presidential election forecasts, which we will discuss in Sec. 10.

9 GENERAL DISCUSSION

9.1 Trust over time

We note that our trust results align with but are slightly different from those Yang et al. reported for the 2022 U.S. midterm elections [113]. They suggest intervals engender the highest trust while dotplot increases trust after the outcome is known. Differences in study design (e.g., viewer environments and forecast distributions) and measures may partly explain the discrepancies. However, the significant difference is that the present work accumulates trust changes over multiple election cycles, and both our results and theirs similarly suggest that dotplot is more robust to maintaining trust over time among the visualizations tested.

9.2 The null-ish effects of probability correction and visual calibration

In designing the experiments, we had hoped that probability correction and/or calibration could dampen the impression that the forecast was “wrong”. However, both techniques show small-to-null or non-conclusive effects on average in our experiments. One reason might be that the voting task and incentives only reward a correct prediction, not considering that a “wrong” forecast could be the same quality as a “correct” forecast. Probability correction was slightly effective in improving people’s attitudinal trust, suggesting that they indeed changed people’s subjective probability to some

extent. Calibration sometimes worsens people’s attitudinal trust, perhaps reminding them of the “mistake”. From the results of our current experiments, we would assume at least that these two techniques do not strongly undermine trust. We speculate that calibration might be more effective when judging forecast quality, and may display a non-linear relationship to the forecast quality. Examining these effects requires a different experimental design that varies forecast quality, which is beyond the scope of the present work.

9.3 Perception and action

It is notable that in our experiments, participants’ perceptions and actions are not always the same, and signals in their actions, or say, differences in their *decisions* [46, 69] are much stronger than those in their attitudes. Perception has been a long-standing interest in both visualization and political science. While perception provides intricate results as mediators between causes and effects, decisions and actions generate more impactful behavioral outcomes. Coming back to the example at the beginning of this manuscript, we may tell our friends that a weather forecaster is unreliable, but do we choose a different forecaster or ignore the next rain forecast? We may express certain views or opinions, but our actions often offer a more genuine reflection of our beliefs and trust. In essence, it is not merely about what we say or think, but more crucially, about what we choose to do in response, and how visualizations and techniques affect or help with such decisions.

9.4 Design for visualization in practice

We also aimed to address the difficulties of applying visualizations in practice. We had to adapt previous work (e.g., [50, 77, 113, 114]) to suit them for the U.S. presidential elections. For example, to apply subjective probability correction, a demonstrated technique, we conducted new experiments, fitted new models, and iterated several times to finally find a design that could be ethically applied in practice.

Also, the two major U.S. political parties are polarized at the elite level [59]. At the time of this research, individuals with higher levels of education tended to lean towards the Democratic party [1], which may partly explain the observed differences in trust. Furthermore, the observed partisan differences could also be partly explained by different individual traits (e.g., spatial memory [97]), working memory capacities [33], visualization literacy [18], or trust in different sciences (e.g., liberals were more trusting of impact scientists and conservatives more trusting of product scientists) [92, 95]. Or that different parties choose to receive information from different sources, impacting the familiarity of different party groups with different types of visualizations. Media outlets like FiveThirtyEight and The New York Times heavily use graphic presentations and have more left-leaning viewers, while RealClearPolitics, mostly relying on text and tables, has more right-leaning readers. This gap between academic research and real-world applications underscores the importance of iterative testing and refinement when translating theoretical insights into practical solutions.

9.5 Limitations and generalization

Our findings primarily apply to U.S. presidential election forecasts, as our experiments were designed specifically for this context. In

Table 1: Participants' voting strategies(498 total responses)

Axis	%	Code	Quote
Use the information provided	34%	probability/odds	<i>I voted if their was a more than 50% chance of wining</i>
	19%	generic answer	<i>I used the website data</i>
	15%	forecasted winner	<i>Based on the statistics of who was projected to win</i>
	8%	margin of victory	<i>Using certain charts to see who had a higher lead in votes</i>
	5%	educational guess	<i>After looking at everything chose my gut feeling</i>
	5%	state information	<i>I used the map with the states to gauge my choice</i>
	4%	risk/payoff	<i>Balancing perceived risk and perceived reward</i>
	2%	trustworthiness	<i>... how trustworthy the site i picked was.</i>
	1%	trend	<i>... I tired to see if there was a trend.</i>

Irrelevant to forecast	13%	always vote	<i>I always voted. It's my responsibility as a US citizen.</i>
	3%	always abstain	<i>Since I was other, I abstained on all votes.</i>
	2%	prior knowledge	<i>Was it typically a red state or blue state?</i>
	<1%	randomly	<i>It was a random decision</i>
...	
Unable to code	2%	-	<i>highest 1 out of 100 ratio</i>

Table 2: What affected participants choosing/trusting a website (498 total responses)

Axis	%	Code	Quote
More task-focused	4%	efficiency	<i>I tried to use the most efficient forecast website</i>
	4%	helpfulness	<i>The one I used was more helpful</i>
	5%	reliability	<i>Which one I felt was most reliable</i>
	22%	accuracy	<i>I chose the one which seemed most accurate when I used it</i>
	19%	understandability	<i>The one that I can understand better</i>
	11%	readability	<i>The balls were easier to read</i>
	7%	clarity	<i>Whichever presented the data most clearly.</i>
	15%	preference	<i>I picked the one I liked</i>
	3%	perception/intuition	<i>I chose the seemingly proficient one</i>
	3%	visual method	<i>I like graphs</i>
	3%	simplicity	<i>I generally used the first one with simple numbers.</i>
	3%	aesthetics	<i>I liked the dots better most visually pleasing</i>
	2%	presentation	<i>I liked the layout of one over the other</i>
	1%	special interests	<i>... because i could watch how people voted</i>
	<1%	for fun	<i>I thought the first one was fun</i>
Less task-focused	<1%	partisan bias	<i>The one that said Roberts would win</i>

No strategy/randomly	6%	-	<i>I randomly picked</i>
Unable to code	6%	-	<i>Number differences</i>

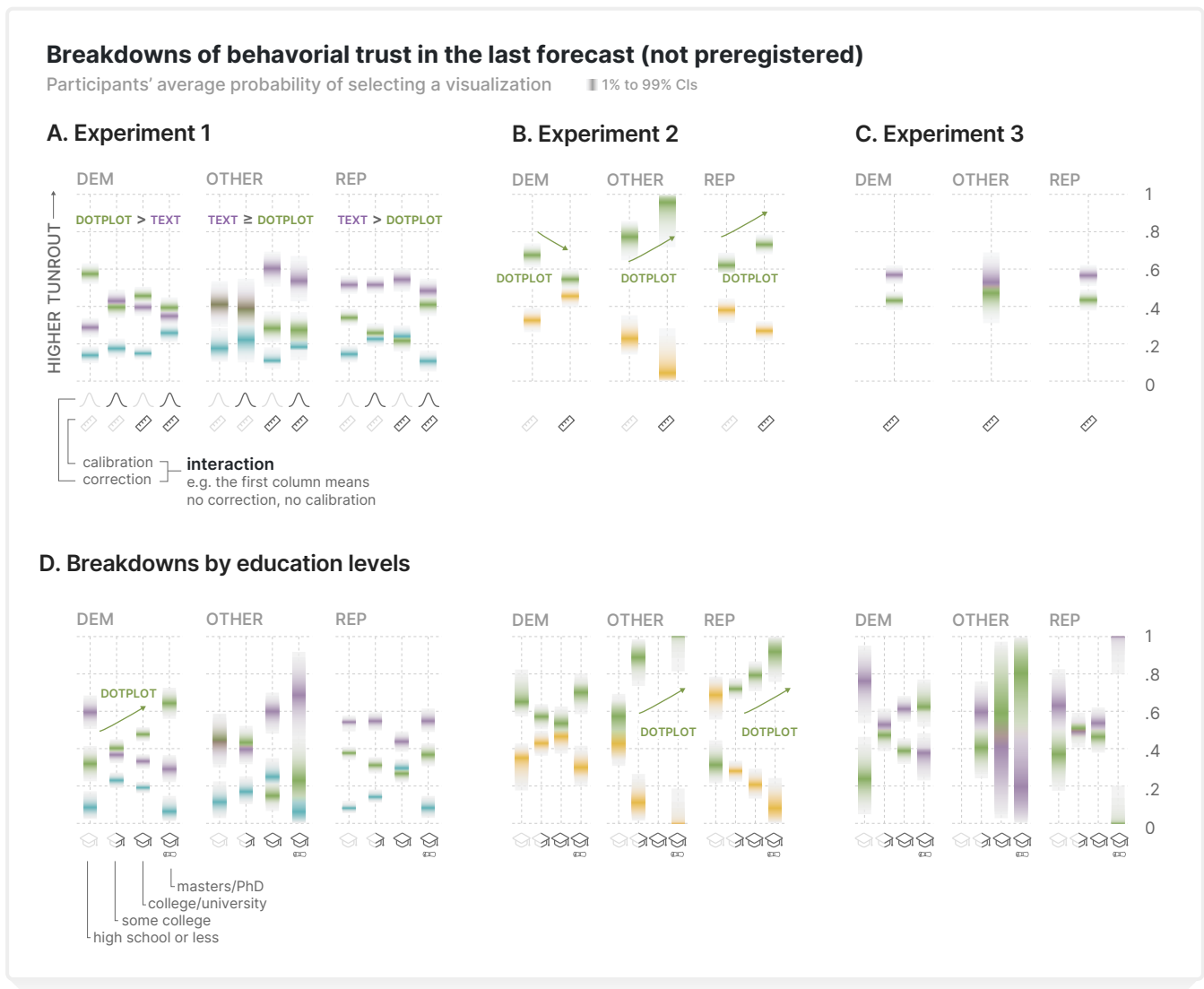


Figure 10: Non-preregistered exploratory analysis for behavioral trust based on partisanship and education levels.

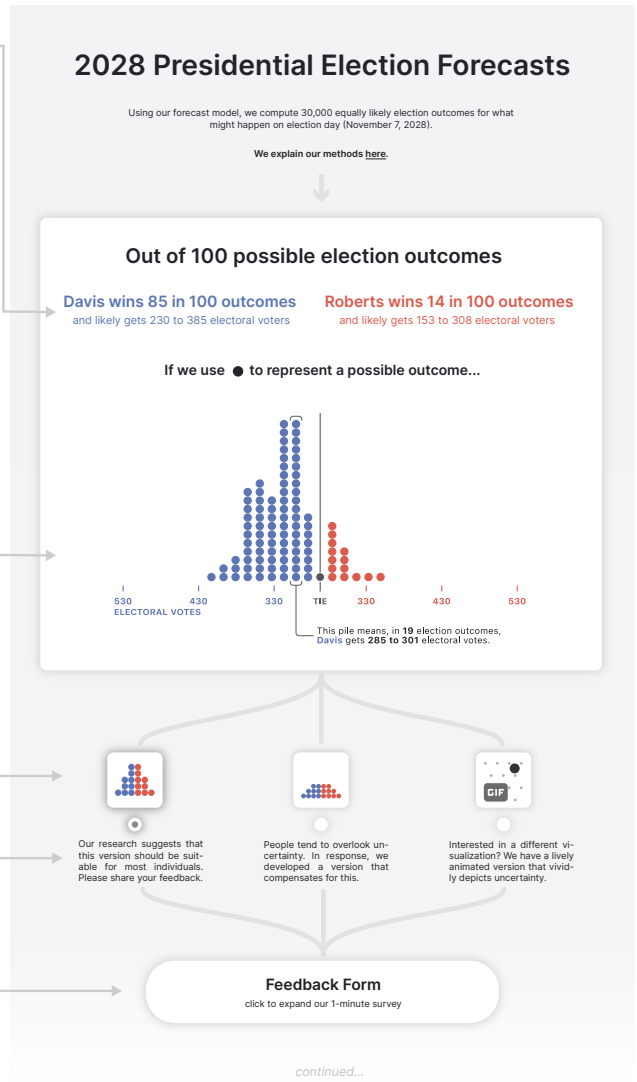
countries with different electoral systems (e.g., multi-party proportional voting systems without an electoral college), adapting our two-party design system may be challenging. Some partisan effects may also not generalize beyond the U.S. political landscape. Also, it is important to note that controlled experiments, while informative, offer only an approximation of the complex and evolving nature of U.S. elections and forecasts. Participants have much to digest in a relatively short period, and our design did not fully capture the gradual shifts in public trust influenced by dynamics in political and media landscapes. Despite these limitations, our experiments highlight the significant potential shifts in trust following an election. Our results of different uncertain displays, participants' desire for clarity and understandability, and the individual differences observed offer insights into maintaining trust over time for communicating uncertainty in general.

10 RECOMMENDATION AND CONCLUSION

Beginning with simulating forecasts of a hypothetical U.S. presidential election, we conduct three experiments to refine our recommendations for uncertainty displays that aim at maintaining trust in election forecasts. Our results suggest that text engenders the highest trust over ten election cycles, especially among participants who support Republicans and Independents; dotplot is most trusted by those supporting Democrats, and in some cases, by participants with advanced education.

Assuming a partisan-neutral design goal—

1. We recommend a mixed layout that can engender the highest trust among the majority of viewers. Start with text summaries; for those viewers adept at interpreting visuals, invite them to the visual—a quantile dotplot.
2. This mixed layout is substantially similar to FiveThirtyEight’s presentation for their 2020 and 2022 forecasts [3, 8] but differs from their 2016 presentation, which used text and a U.S. map [2]. Nonetheless, some individuals may prefer different designs, and forecasters may want to test alternative designs, possibly integrating more options into the website interface.
3. Responsible forecasters may consider gauging their viewers’ trust level. While our approach involves a sequence of experiments, in a real-world setting, the choice between display options can be modeled as a multi-armed bandit problem [15]. Each display would be an arm, and the optimal display to show could be determined using Thompson sampling (e.g., [15, 34], see [42] for an example in HCI). Forecasters may also embed a short survey to obtain demographics and partisanship.
4. Our qualitative results highlight the need for forecast presentations to prioritize understandability, readability, clarity, and accuracy. The educational background of viewers may significantly influence their trust and preferences. Forecasters might consider simplifying the interface to cater to a broader audience. Additionally, we balanced the two parties when analyzing voter turnout and considered our turnout results to be specific to the current experiments. Future studies might explore differential turnouts in more realistic settings (e.g., a field experiment [58]); for instance, certain displays might unintentionally motivate one party over another to vote. Finally, we also pose unresolved questions regarding probability correction, visual calibration, and state-level presentations. When constructing the interface, we found a contrast between before and after election day helpful for checking data (e.g., the snake chart). While our current results do not provide substantial evidence for or against these attempts, we hope our insights will spur future research, especially for those being planned for the 2024 U.S. presidential election.



ACKNOWLEDGMENTS

This research is supported by NSF 2127309 to the Computing Research Association for the CIFellows Project. The authors thank Maryam Hedayati for her help with the experiments, and Mandi Cai, Abhraneel Sarma, Lily W. Ge for their feedback on the manuscript. The authors also thank Angelos Chatzimpampas, Yuan “Charles” Cui, Ziyang Guo, Priyanka Nanayakkara, Hyeok Kim, Jessica Hullman, and Haochen Zhang for their valuable feedback on the research. Icons are designed by the authors or under a Flaticon license.

REFERENCES

- [1] 2016. How Educational Differences Are Widening America’s Political Rift. <https://www.nytimes.com/2021/09/08/us/politics/how-college-graduates-vote.html>.
- [2] 2016. Who Will Win the Presidency? <https://projects.fivethirtyeight.com/2016-election-forecast/>.
- [3] 2020. 2020 Election Forecast. <https://projects.fivethirtyeight.com/2020-election-forecast/senate/>.
- [4] 2020. 2020 Senate Forecast. <https://projects.fivethirtyeight.com/2020-election-forecast/>.
- [5] 2020. Forecasting the U.S. elections. <https://projects.economist.com/us-2020-forecast/president>.
- [6] 2020. Voters Rarely Switch Parties, but Recent Shifts Further Educational, Racial Divergence. <https://www.pewresearch.org/politics/2020/08/04/voters-rarely-switch-parties-but-recent-shifts-further-educational-racial-divergence/>.
- [7] 2021. The Economist’s German Election Model Forecasts Three Possible Governing Coalitions. <https://www.economistgroup.com/group-news/the-economist-the-economists-german-election-model-forecasts-three-possible-governing>.
- [8] 2022. 2022 Election Forecast. <https://projects.fivethirtyeight.com/2022-election-forecast/>.
- [9] 2022. The race for the Senate is very close. <https://www.economist.com/interactive/us-midterms-2022/forecast/senate>.

- [10] 2023. How Good Are FiveThirtyEight Forecasts? <https://projects.fivethirtyeight.com/checking-our-work>.
- [11] 2023. Metaculus Track Record. <https://www.metaculus.com/questions/track-record/>.
- [12] 2023. Voter Turnout, 2018–2022. <https://www.pewresearch.org/politics/2023/07/12/voter-turnout-2018-2022/>.
- [13] Alan I. Abramowitz and Steven Webster. 2016. The Rise of Negative Partisanship and the Nationalization of U.S. Elections in the 21st Century. *Electoral Studies* 41 (2016), 12–22. <https://doi.org/10.1016/j.electstud.2015.11.001>
- [14] Marina Agranov, Jacob K Goeree, Julian Romero, and Leeat Yariv. 2017. What Makes Voters Turn Out: The Effects of Polls and Beliefs. *Journal of the European Economic Association* 16, 3 (08 2017), 825–856. <https://doi.org/10.1093/jeea/jvx023>
- [15] Shipra Agrawal and Navin Goyal. 2012. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the Annual Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 23)*, Shie Mannor, Nathan Srebro, and Robert C. Williamson (Eds.). PMLR, Edinburgh, Scotland, 39.1–39.26.
- [16] Ali M. Ahmed and Osvaldo Salas. 2009. The Relationship between Behavioral and Attitudinal Trust: A Cross-cultural Study. *Review of Social Economy* 67, 4 (2009), 457–482. <https://doi.org/10.1080/00346760902908625>
- [17] Stephen Ansolabehere and Shanto Iyengar. 1994. Of Horseshoes and Horse Races: Experimental Studies of the Impact of Poll Results on Electoral Behavior. *Political Communication* 11, 4 (1994), 413–430. <https://doi.org/10.1080/10584609.1994.9963048>
- [18] Maria Avgerinou and John Ericson. 1997. A Review of the Concept of Visual Literacy. *British Journal of Educational Technology* 28, 4 (1997), 280–291. <https://doi.org/10.1111/1467-8535.00035>
- [19] Andrea Ballatore, David Gordon, and Alexander P. Boone. 2019. Sonifying Data Uncertainty with Sound Dimensions. *Cartography and Geographic Information Science* 46, 5 (2019), 385–400. <https://doi.org/10.1080/15230406.2018.1495103>
- [20] Melanie Bancillon, Zhengliang Liu, and Alvitia Ottley. 2020. Let's Gamble: How a Poor Visualization Can Elicit Risky Behavior. In *The IEEE Visualization Conference, short paper*. 196–200. <https://doi.org/10.1109/VIS47514.2020.00046>
- [21] Matthew Barnfield. 2020. Think Twice before Jumping on the Bandwagon: Clarifying Concepts in Research on the Bandwagon Effect. *Political Studies Review* 18, 4 (2020), 553–574. <https://doi.org/10.1177/1478929919870691>
- [22] Alejandro Barredo Arrieta, Natalia Diaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbedo, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [23] Marco Battaglini, Rebecca B. Morton, and Thomas R. Palfrey. 2010. The Swing Voter's Curse in the Laboratory. *The Review of Economic Studies* 77, 1 (01 2010), 61–89. <https://doi.org/10.1111/j.1467-937X.2009.00569.x>
- [24] Donald W Beachler, Matthew L Bergbower, Chris Cooper, David F Damore, Bas Van Dooren, Sean D Foreman, Rebecca D Gill, Henriët Hendriks, Donna Hoffmann, Rafael Jacob, et al. 2015. *Presidential Swing States: Why Only Ten Matter*. Lexington Books.
- [25] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. 2016. On the Relevance of Irrelevant Alternatives. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 963–973. <https://doi.org/10.1145/2872427.2883025>
- [26] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, 1 (1995), 122–142. <https://doi.org/10.1006/game.1995.1027>
- [27] André Blais. 2000. *To Vote or Not to Vote?: The Merits and Limits of Rational Choice Theory*. University of Pittsburgh Pre.
- [28] Olivier Blanchard, Christopher G. Collins, Mohammad R. Jahan-Parvar, Thomas Pellet, and Beth Anne Wilson. 2018. Why Has the Stock Market Risen So Much Since the U.S. Presidential Election? *FRB International Finance Discussion Paper* (2018). <https://doi.org/10.17016/IFDP.2018.1235>
- [29] Benjamin T. Blankenship and Abigail J. Stewart. 2019. Threat, trust, and Trump: identity and voting in the 2016 presidential election. *Politics, Groups, and Identities* 7, 3 (2019), 724–736. <https://doi.org/10.1080/21565503.2019.1633932>
- [30] James Dean Brown. 2011. Likert Items and Scales of Measurement. *Statistics* 15, 1 (2011), 10–14.
- [31] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the International Conference on Intelligent User Interfaces*. 454–464. <https://doi.org/10.1145/3377325.3377498>
- [32] Paul-Christian Bürkner, Jonah Gabry, Matthew Kay, and Aki Vehtari. 2023. posterior: Tools for Working with Posterior Distributions. <https://mc-stan.org/posterior/> R package version 1.4.1.9000.
- [33] Spencer C. Castro, P. Samuel Quinan, Helia Hosseinpour, and Lace Padilla. 2022. Examining Effort in 1D Uncertainty Communication Using Individual Differences in Working Memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 411–421. <https://doi.org/10.1109/TVCG.2021.3114803>
- [34] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Proceedings of the International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2249–2257.
- [35] A. Chatzimpampas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* 39 (2020), 713–756. <https://doi.org/10.1111/cgf.14034>
- [36] Sungeun Chung, Yu-Jin Heo, and Jung-Hyun Moon. 2017. Perceived Versus Actual Polling Effects: Biases in Perceptions of Election Poll Effects on Candidate Evaluations. *International Journal of Public Opinion Research* 30, 3 (2017), 420–442. <https://doi.org/10.1093/ijpor/edx004>
- [37] Michael Correll and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE TVCG* 20, 12 (2014), 2142–2151. <https://doi.org/10.1109/TVCG.2014.2346298>
- [38] Fred Cutler, J Scott Matthews, and Mark Pickup. [n. d.]. Could Polls Matter? Evaluating the Preconditions for Poll Effects. ([n. d.]).
- [39] Jens Olav Dahlgaard, Jonas Hedegaard Hansen, Kasper M. Hansen, and Martin V. Larsen. 2016. How are Voters Influenced by Opinion Polls? The Effect of Polls on Voting Behavior and Party Sympathy. *World Political Science* 12, 2 (2016), 283–300. <https://doi.org/doi:10.1515/wps-2016-0012>
- [40] Jens Olav Dahlgaard, Jonas Hedegaard Hansen, Kasper M. Hansen, and Martin V. Larsen. 2017. How Election Polls Shape Voting Behaviour. *Scandinavian Political Studies* 40, 3 (2017), 330–343. <https://doi.org/10.1111/1467-9477.12094>
- [41] T. K. Das and Bing-Sheng Teng. 2004. The Risk-Based View of Trust: A Conceptual Framework. *Journal of Business and Psychology* 19, 1 (2004), 85–116. <https://doi.org/10.1023/B:JOBU.0000040274.23551.1b>
- [42] Nediya Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. <https://doi.org/10.1145/3313831.3376584>
- [43] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics* 12, 2 (2020), 459–478. <https://doi.org/10.1007/s12369-019-00596-x>
- [44] Delia Diaconășu, Seyed Mehdiian, and Ovidiu Stoica. 2023. The Global Stock Market Reactions to the 2016 U.S. Presidential Election. *SAGE Open* 13, 2 (2023). <https://doi.org/10.1177/21582440231181352>
- [45] Nicholas Diakopoulos. 2022. Predictive journalism: On the role of computational prospection in news media. *Tow Center for Digital Journalism* (2022).
- [46] Evanthis Dimara and John Stasko. 2022. A Critical Reflection on Visualization Research: Where Do Decision Making Tasks Hide? *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1128–1138. <https://doi.org/10.1109/TVCG.2021.3114813>
- [47] Hamza Elhamedadi, Aimen Gaba, Yea-Seul Kim, and Cindy Xiong. 2022. How Do We Measure Trust in Visual Data Communication?. In *IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 85–92. <https://doi.org/10.1109/BELIV57783.2022.00014>
- [48] Hamza Elhamedadi, Adam Stefkovics, Johanna Beyer, Eric Moerth, Hanspeter Pfister, Cindy Xiong Bearfield, and Carolina Nobre. 2023. Vistrust: a Multidimensional Framework and Empirical Study of Trust in Data Visualizations. , 11 pages. <https://doi.org/10.1109/TVCG.2023.3326579>
- [49] Mike Farjam. 2020. The Bandwagon Effect in an Online Voting Experiment With Real Political Organizations. *International Journal of Public Opinion Research* 33, 2 (06 2020), 412–421. <https://doi.org/10.1093/ijpor/edaa008>
- [50] 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 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3173574.3173718>
- [51] Anthony Fowler. 2015. Regular Voters, Marginal Voters and the Electoral Effects of Turnout. *Political Science Research and Methods* 3, 2 (2015). <https://doi.org/10.1017/psrm.2015.18>
- [52] Thomas Fujiwara, Kyle Meng, and Tom Vogl. 2016. Habit Formation in Voting: Evidence from Rainy Elections. *American Economic Journal: Applied Economics* 8, 4 (October 2016), 160–88. <https://doi.org/10.1257/app.20140533>
- [53] Aimen Gaba, Zhanna Kaufman, Jason Cheung, Marie Shvake, Kyle Wm. Hall, Yuriy Brun, and Cindy Xiong Bearfield. 2023. My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning. In *Proceedings of the IEEE Visualization and Visual Analytics Conference*.
- [54] Jonah Gabry and Rok Češnovar. 2023. CmdStanR: the R interface to CmdStan. <https://mc-stan.org/users/interfaces/cmdstan>
- [55] Francis Galton. 1889. *Natural Inheritance*. Vol. 42. Macmillan.
- [56] Andrew Gelman, Jessica Hullman, Christopher Wlezien, and George Elliott Morris. 2020. Information, Incentives, and Goals in Election Forecasts. *Judgment and Decision Making* 15, 5 (2020), 863–880. <https://doi.org/10.1017/>

- S1930297500007981
- [57] Alan Gerber, Mitchell Hoffman, John Morgan, and Collin Raymond. 2020. One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout. *American Economic Journal: Applied Economics* 12, 3 (2020), 287–325. <https://doi.org/10.1257/app.20180574>
- [58] ALAN S. GERBER, DONALD P. GREEN, and CHRISTOPHER W. LARIMER. 2008. Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review* 102, 1 (2008), 33–48. <https://doi.org/10.1017/S000305540808009X>
- [59] Daniel Q. Gillion, Jonathan M. Ladd, and Marc Meredith. 2020. Party Polarization, Ideological Sorting and the Emergence of the US Partisan Gender Gap. *British Journal of Political Science* 50, 4 (2020), 1217–1243. <https://doi.org/10.1017/S0007123418000285>
- [60] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- [61] Miriam Greis, Aditi Joshi, Ken Singer, Albrecht Schmidt, and Tonja Machulla. 2018. Uncertainty Visualization Influences How Humans Aggregate Discrepant Information. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–12. <https://doi.org/10.1145/3173574.3174079>
- [62] 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. <https://doi.org/10.2307/234419>
- [63] Spencer E. Harpe. 2015. How to Analyze Likert and Other Rating Scale Data. *Currents in Pharmacy Teaching and Learning* 7, 6 (2015), 836–850. <https://doi.org/10.1016/j.cptl.2015.08.001>
- [64] Merlin Heidemanns, Andrew Gelman, and G. Elliott Morris. 2020. An Updated Dynamic Bayesian Forecasting Model for the U.S. Presidential Election. *Harvard Data Science Review* 2, 4 (2020). <https://doi.org/10.1162/99608f92.fc62f1e1>
- [65] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besançon. 2021. Can Visualization Alleviate Dichotomous Thinking? Effects of Visual Representations on the Cliff Effect. *IEEE TVCG* 27, 8 (2021), 3397–3409. <https://doi.org/10.1109/TVCG.2021.3073466>
- [66] D. Sunshine Hillygus. 2011. The Evolution of Election Polling in the United States. *Public Opinion Quarterly* 75, 5 (12 2011), 962–981. <https://doi.org/10.1093/poq/nfr054>
- [67] Eli Holder and Cindy Xiong Bearfield. 2023. Polarizing Political Polls: How Visualization Design Choices Can Shape Public Opinion and Increase Political Polarization. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11. <https://doi.org/10.1109/TVCG.2023.3326512>
- [68] Jessica Hullman. 2020. Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 130–139. <https://doi.org/10.1109/TVCG.2019.2934287>
- [69] Jessica Hullman, Xiaoqi Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 903–913. <https://doi.org/10.1109/TVCG.2018.2864889>
- [70] 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 (11 2015), 1–25. <https://doi.org/10.1371/journal.pone.0142444>
- [71] Kim Fridkin Kahn. 1992. Does Being Male Help? An Investigation of the Effects of Candidate Gender and Campaign Coverage on Evaluations of U.S. Senate Candidates. *The Journal of Politics* 54, 2 (1992), 497–517. <https://doi.org/10.2307/2132036>
- [72] 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>
- [73] Ben Kaplan. 2019. Experimental Electionomics: How Election Forecasts Influence Voter Turnout. Bachelors dissertation.
- [74] Matthew Kay. 2023. *ggblend: Blending and Compositing Algebra for ggplot2*. <https://doi.org/10.5281/zenodo.7963886> R package version 0.1.0.
- [75] Matthew Kay. 2023. *tidybayes: Tidy Data and Geoms for Bayesian Models*. <https://doi.org/10.5281/zenodo.1308151>
- [76] Matthew Kay. 2023. Visualizations of Distributions and Uncertainty in the Grammar of Graphics. In *Proceedings of the IEEE Visualization and Visual Analytics Conference*.
- [77] 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 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/2858036.2858558>
- [78] Ryan Kennedy, Stefan Wojcik, and David Lazer. 2017. Improving Election Prediction Internationally. *Science* 355, 6324 (2017), 515–520. <https://doi.org/10.1126/science.aal2887>
- [79] Shahedul Huq Khandkar. 2009. Open Coding. *University of Calgary* 23 (2009), 2009.
- [80] William C Kimberling. 1992. *The electoral college*. Vol. 1. National Clearinghouse on Election Administration, Federal Election Commission.
- [81] Peter Knapp, Peter Gardner, Brian McMillan, David K Raynor, and Elizabeth Woolf. 2012. Evaluating a Combined (Frequency and Percentage) Risk Expression to Communicate Information on Medicine Side Effects to Patients. *International Journal of Pharmacy Practice* 21, 4 (11 2012), 226–232. <https://doi.org/10.1111/j.2042-7174.2012.00254.x>
- [82] Morgane Koval and Yvonne Jansen. 2022. Do You See What You Mean? Using Predictive Visualizations to Reduce Optimism in Duration Estimates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 30, 19 pages. <https://doi.org/10.1145/3491102.3502010>
- [83] Richard R. Lau and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. Cambridge University Press.
- [84] Pamela Lenton and Paul Mosley. 2011. Incentivising trust. *Journal of Economic Psychology* 32, 5 (2011), 890–897. <https://doi.org/10.1016/j.joep.2011.07.005>
- [85] David K. Levine and Thomas R. Palfrey. 2007. The Paradox of Voter Participation? A Laboratory Study. *American Political Science Review* 101, 1 (2007), 143–158. <https://doi.org/10.1017/S0003055407070013>
- [86] Michael S. Lewis-Beck. 1985. Election Forecasts in 1984: How Accurate Were They? *PS: Political Science & Politics* 18, 1 (1985), 53–62. <https://doi.org/10.2307/418806>
- [87] Le Liu, Alexander P. Boone, Ian T. Ruginski, Lace Padilla, Mary Hegarty, Sarah H. Creem-Regehr, William B. Thompson, Cem Yuksel, and Donald H. House. 2017. Uncertainty Visualization by Representative Sampling from Prediction Ensembles. *IEEE TVCG* 23, 9 (2017), 2165–2178. <https://doi.org/10.1109/TVCG.2016.2607204>
- [88] Mark J. Machina and David Schmeidler. 1992. A More Robust Definition of Subjective Probability. *Econometrica* 60, 4 (1992), 745–780. <https://doi.org/10.2307/2951565>
- [89] Maria Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *Australasian Conference on Information Systems*, Vol. 53. 6–8.
- [90] Louis Sandy Maisel. 2022. *American political parties and elections: A very short introduction*. Vol. 169. Oxford University Press.
- [91] Sandy L Maisel and Mark D Brewer. 2009. *Parties and elections in America: The electoral process*. Rowman & Littlefield.
- [92] Aaron M McCright, Katherine Dentzman, Meghan Charters, and Thomas Dietz. 2013. The Influence of Political Ideology on Trust in Science. *Environmental Research Letters* 8, 4 (2013), 044029.
- [93] Monika L. Mcdermott. 1998. Race and Gender Cues in Low-Information Elections. *Political Research Quarterly* 51, 4 (1998), 895–918. <https://doi.org/10.1177/106591299805100403>
- [94] John M. McGuirl and Nadine B. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors* 48, 4 (2006), 656–665. <https://doi.org/10.1518/001872006779166334> PMID: 17240714.
- [95] Erik C. Nisbet, Kathryn E. Cooper, and R. Kelly Garrett. 2015. The Partisan Brain: How Dissonant Science Messages Lead Conservatives and Liberals to (Dis)Trust Science. *The ANNALS of the American Academy of Political and Social Science* 658, 1 (2015), 36–66. <https://doi.org/10.1177/0002716214555474>
- [96] Andreas Ortmann, John Fitzgerald, and Carl Boeing. 2000. Trust, Reciprocity, and Social History: A Re-examination. *Experimental Economics* 3, 1 (2000), 81–100. <https://doi.org/10.1023/A:1009946125005>
- [97] Alivitta Otlely, Evan M. Peck, Lane T. Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A. Taylor, Paul K. J. Han, and Remco Chang. 2016. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 529–538. <https://doi.org/10.1109/TVCG.2015.2467758>
- [98] Lace Padilla, Sarah Dryhurst, Helia Hosseinpour, and Andrew Kruczkiewicz. 2021. Multiple hazard uncertainty visualization challenges and paths forward. *Frontiers in Psychology* (2021). <https://doi.org/10.3389/fpsyg.2021.579207>
- [99] Lace Padilla, Racquel Fygenon, Spencer C. Castro, and Enrico Bertini. 2023. Multiple Forecast Visualizations (MFVs): Trade-offs in Trust and Performance in Multiple COVID-19 Forecast Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 12–22. <https://doi.org/10.1109/TVCG.2022.3209457>
- [100] Ellen Peters, P. Sol Hart, and Liana Fraenkel. 2011. Informing Patients: The Influence of Numeracy, Framing, and Format of Side Effect Information on Risk Perceptions. *Medical Decision Making* 31, 3 (2011), 432–436. <https://doi.org/10.1177/0272989X10391672> PMID: 21191122.
- [101] José Pinheiro and Douglas Bates. 2006. *Mixed-Effects Models in S and S-PLUS*. Springer science & business media.
- [102] Daron R. Shaw. 2008. *The race to 270: The electoral college and the campaign strategies of 2000 and 2004*. University of Chicago Press.
- [103] John Sides, Michael Tesler, and Lynn Vavreck. 2017. The 2016 US election: How Trump lost and won. *Journal of Democracy* 28, 2 (2017), 34–44. <https://doi.org/10.1353/jod.2017.0022>
- [104] Stan Development Team. 2023. RStan: the R interface to Stan. <http://mc-stan.org/> R package version 2.21.5.

- [105] Sjoerd B. Stolwijk, Andreas R. T. Schuck, and Claes H. de Vreese. 2016. How Anxiety and Enthusiasm Help Explain the Bandwagon Effect. *International Journal of Public Opinion Research* 29, 4 (2016), 554–574. <https://doi.org/10.1093/ijpor/edw018>
- [106] Tom W. G. van der Meer, Armen Hakhverdian, and Loes Aaldering. 2015. Off the Fence, Onto the Bandwagon? A Large-Scale Survey Experiment on Effect of Real-Life Poll Outcomes on Subsequent Vote Intentions. *International Journal of Public Opinion Research* 28, 1 (02 2015), 46–72. <https://doi.org/10.1093/ijpor/edu041>
- [107] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-Normalization, Folding, and Localization: An Improved Rhat for Assessing Convergence of MCMC (with discussion). *Bayesian Analysis* (2021).
- [108] 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>
- [109] Rick K. Wilson and Catherine C. Eckel. 2011. Innovation and Intellectual Property Rights. In *Cambridge handbook of experimental political science*, James N Druckman, Donald P Greene, and James H Kuklinski (Eds.). Cambridge University Press, Chapter 17, 243–257.
- [110] Benedict Witzembergera and Nicholas Diakopoulos. 2023. Election Predictions in the News: How Users Perceive and Respond to Visual Election Forecasts. *Information, Communication & Society* (2023), 1–22. <https://doi.org/10.1080/1369118X.2023.2230267>
- [111] Cindy Xiong, Ali Sarvghad, Daniel G. Goldstein, Jake M. Hofman, and Gagatay Demiralp. 2022. Investigating Perceptual Biases in Icon Arrays. In *ACM CHI*. Article 137, 12 pages. <https://doi.org/10.1145/3491102.3501874>
- [112] Toshio Yamagishi, Satoshi Akutsu, Kisuk Cho, Yumi Inoue, Yang Li, and Yoshie Matsumoto. 2015. Two-Component Model of General Trust: Predicting Behavioral Trust from Attitudinal Trust. *Social Cognition* 33, 5 (2015), 436–458. <https://doi.org/10.1521/soco.2015.33.5.436>
- [113] Fumeng Yang, Mandi Cai, Chloe Mortenson, Hoda Fakhari, Ayse Lokmanoglu, Jessica Hullman, Steven Franconeri, Nicholas Diakopoulos, Erik C. Nisbet, and Matthew Kay. 2023. Swaying the Public? Impacts of Election Forecast Visualizations on Emotion, Trust, and Intention in the 2022 U.S. Midterms. In *Proceedings of the IEEE Visualization and Visual Analytics Conference*.
- [114] Fumeng Yang, Maryam Hedayati, and Matthew Kay. 2023. Subjective Probability Correction for Uncertainty Representations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 17 pages. <https://doi.org/10.1145/3544548.3580998>
- [115] Michaël Zamo and Philippe Naveau. 2018. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences* 50, 2 (2018), 209–234. <https://doi.org/10.1007/s11004-017-9709-7>
- [116] Dongping Zhang, Eytan Adar, and Jessica Hullman. 2022. Visualizing Uncertainty in Probabilistic Graphs with Network Hypothetical Outcome Plots (NetHOPs). *IEEE TVCG* 28, 1 (2022), 443–453. <https://doi.org/10.1109/TVCG.2021.3114679>
- [117] 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>