

**High Chances and Close Margins:
How Equivalent Forecasts Yield Different Beliefs**

Oleg Urminsky (oleg.urminsky@chicagobooth.edu)
University of Chicago
5807 S. Woodlawn Ave, Chicago, IL 60637 USA

Luxi Shen (luxi.shen@cuhk.edu.hk)
Chinese University of Hong Kong
12 Chak Cheung Street, Hong Kong

First Draft: 2/8/2018
Current Draft: 2/13/2020

Please contact authors for current version before citing.

Abstract

Statistical forecasts are increasingly prevalent. How do forecasts affect people's beliefs about corresponding future events? This research proposes that the format in which the forecast is communicated biases its interpretation. We contrast two common forecast formats: *chance* (the forecasted probability that an outcome will occur; e.g., the likelihood that a political candidate or a sports team will win) versus *margin* (the forecasted amount by which an outcome will occur; e.g., by how many points the favored political candidate or sports team will win). Across six studies (total N = 2,995; plus 12 replication and generalization studies with an additional total N = 3,459), we document a robust chance-margin discrepancy: chance forecasts lead to more extreme beliefs about outcome occurrences than do margin forecasts. This discrepancy persists over time in the interpretation of publicly available forecasts about real-world events (e.g., the 2016 U.S. presidential election), replicates even when the forecasts are strictly statistically equivalent, and has downstream consequences for attitudes toward election candidates and sports betting decisions. The findings in this research have important societal implications for how forecasts are communicated and for how people use forecast information to make decisions.

1. Introduction

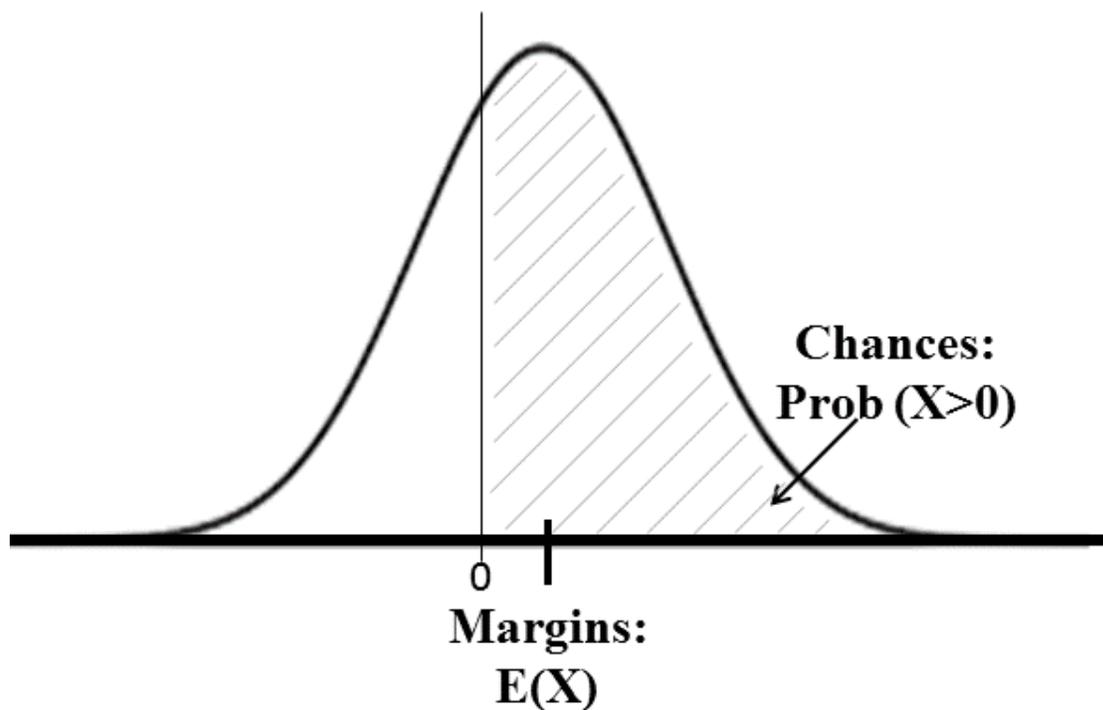
Forecasts are pervasive. People frequently encounter forecasts about weather, stock market performance, sports outcomes, and election results. Election forecasts reported in the mass media, for example, contribute to voters' beliefs about important events such as the U.S. presidential election and the U.K. Brexit referendum. In the summer of 2016, the media described the status of pre-election opinions with headlines that communicated forecasts in different ways, such as “The Upshot's Election Forecast: 76% Chance of a Clinton Victory” in the New York Times (Katz, 2016) and “Hillary Clinton Maintains 5-Point Lead Over Donald Trump” in the Wall Street Journal (Hook, 2016). Such forecasts direct the readers' attention to either the probability of a candidate winning or the predicted difference in the votes to be received by each candidate. While researchers have debated the merits of specific forecast formats (e.g., probabilistic forecasts: Peachey et al., 2013), it remains unclear how *differences* between forecast formats impact people's beliefs. Thus, our research explores this critical topic by asking: How do people comprehend forecasts in different formats and what are the consequences of those formats on people's beliefs and preferences?

2. The Formation and Formats of Forecasts

Since the early days of statistics, experts have used different approaches to generate and communicate forecasts about future events. For example, in the case of election forecasts, past generations of pundits typically used either their “inside view” personal judgment (Kahnemann & Lovallo, 1993) or exemplars (e.g., “bellwether” locations) to make predictions. A large literature has attempted to develop more scientific practices for generating accurate forecasts (see Armstrong et al., 2015, for a review). Accordingly, forecasters increasingly rely on more

quantitative approaches, such as political polling results and model-based polling aggregation (Hilygus, 2011). These modern approaches first collect data and then use the data to generate a full, estimated probability distribution of the potential outcomes. Mass audiences, however, rarely see these distribution-level predictive statistics (or even proxies, such as confidence ranges). Instead, audiences are typically shown a single statistic that summarizes the full probability distribution. Even forecasts generated for industry experts tend to report a summary statistic (e.g., consensus earnings forecasts: Coën et al., 2009; probability of recession: Torry, 2019).

Figure 1: A chance forecast and a margin forecast that are based on the same prediction distribution.



These forecasts are commonly communicated in one of two formats: (a) the *chance* of a focal outcome (e.g., the predicted probability that a political candidate will win the election or that a sports team will win the game) or (b) the expected *margin* (e.g., the predicted difference in vote shares received by the winning vs. by the losing candidate, or the expected point spread between two teams). As long as the two formats are based on the same underlying probability distribution (Figure 1), they are statistically equivalent—but are they also psychologically equivalent?

3. The Interpretations of Forecasts

Although a substantial body of research has examined how people process numeric information, the impact of forecast format on beliefs is still poorly understood. The extant literature yields diverging predictions about the accuracy with which people process statistical information in a given format.

One possibility is that people do see through the format difference, thus correctly processing chance and margin forecasts as statistically-equivalent statements. This possibility is theoretically grounded in the view, prevalent in cognitive and mathematical psychology, that people are skilled intuitive statisticians, particularly in domains in which they have repeated exposure and ample experience, such as politics and sports. Recent influential research in cognitive psychology has argued that ostensibly systematically biased decision making can actually be explained with Bayesian models that account for uncertainty or the computational cost of complex inferences (Lieder et al., 2012; Pouget et al., 2013).

According to this view, people have a near-optimal ability to represent and update probability distribution information (Griffiths & Tenenbaum, 2006), and this ability may even be

spontaneous at the neural level (Knill & Pouget, 2004; Ma et al., 2006). If this is indeed the case, people should be able to interpret corresponding chance and margin forecasts correctly based on the underlying outcome distribution—thereby extracting equivalent information from the different forecast formats. Nevertheless, a recent debate (Bowers & Davis, 2012; Griffiths et al., 2012; Jones & Love, 2011) has questioned the degree to which normative Bayesian processes provide a realistic process description of how people perform complex reasoning and decision tasks.

Another possibility is that the way in which people process numerical information, including forecasts, may depend on the format in which the information is communicated. This prediction stems from the position that people are poor intuitive statisticians—contrary to the previous view. A classic literature on intuitive statistics suggests that people are often systematically biased in statistical reasoning, particularly when making inferences from samples (Kahneman & Tversky, 1972; Peterson et al., 1968; Wheeler & Beach, 1968). Research has identified comprehension problems specifically in the context of forecasts (e.g., Fischhoff, 1994, Flugstad & Windshitl, 2003). People have also been shown to form different interpretations when equivalent information is framed differently (e.g., “25% fat” vs. “75% lean,” Levin & Gaeth, 1988; see also Kuhberger, 1998, and Tversky & Kahneman, 1981). However, these kinds of framing effects have not been studied in the context of forecasts. That is, the framing literature does not take a theoretical stance on whether people calibrate chance and margin forecasts differently and, if they do, which forecast is interpreted as a more extreme prediction.

Nevertheless, recent work on numerical cognition has demonstrated some important consequences of other information formats on beliefs and decision making. Larrick and Soll (2008) found that although automobile buyers care about gas consumption and carbon emissions,

they misunderstand “miles per gallon” (MPG) information, incorrectly assuming that the amount of gas consumed by a car decreases as a linear function of the car’s MPG. Thus, people may fall short of the Bayesian ideal, processing forecasts with an incomplete understanding that yields mis-calibrated interpretations.

4. Study Overview

We conducted a series of high-powered studies (the majority of which are pre-registered) that test whether people derive the same meaning from equivalent chance and margin forecasts. Across the studies, we identified a general forecast-format bias which is manifested as systematic errors in estimation, attitudes, and preferences. In particular, we scrutinized and documented the *chance-margin discrepancy*, a specific form of forecast-format bias that captures the estimation mismatch between the two formats themselves. We report six studies in this article and another 12 studies in the Appendix.

In the first set of studies (Studies 1–4, as well as Studies A1–A9 in the Appendix), we examined the impact of forecast formats in various real-life contexts for which forecasts are publicly available and widely viewed. In Study 1, we found that viewing a chance forecast that was publicly available during the 2018 U.S. presidential election leads to more extreme attitudes about an actual election than viewing the equivalent margin forecast from the same source. Next, in Study 2, we documented the chance-margin discrepancy, a systematic forecast-format bias when making estimates: chance forecasts over the course of the election consistently yielded overestimated margin forecasts, while margin forecasts during the same time yielded underestimated chance forecasts. We observed and documented the same chance-margin

discrepancy in a different domain—sports outcome forecasts—in Study 3. In Study 4, we extended these findings to incentive-compatible sports betting choices.

In the second set of studies (Studies 5–6, as well as Studies A11 and A12 in the Appendix), we investigated the calibration process and assessed alternative accounts. In Study 5, within the context of genre differences in movie ratings, we provided participants with detailed information about the statistical distribution and replicated the chance-margin discrepancy using novel optimal forecasts. As an even stronger test of internal validity, we elicited incentive-compatible estimates of forecasts in a statistical scenario that ensured the strict mathematical equivalence of the chance and margin forecasts and provided participants with full information (Study 6).

We also collected process measures in Studies 5 and 6, which suggested that people primarily use an intuition-based process to think about forecasts, and people made more accurate estimates when they reported incorporating the provided forecast into their judgments. Nevertheless, we found no evidence that engaging in formal statistical reasoning improves the accuracy of forecast interpretation, or that the chance-margin discrepancy is explained by inaccurate statistical beliefs. We provide additional details and replication studies in the Online Appendix, and all experimental details and data are available at OSF (<https://osf.io/f4ys6/>).

5. Study 1: Forecast Format Impacts Subjective Assessments

During the 2016 U.S. presidential election, we tested whether people reached the same conclusions from equivalent chance and margin forecasts. We presented participants with either a chance forecast (probability of election outcome) or a presumably equivalent margin forecast (predicted vote share) and asked participants to rate their attitude toward the forecast news. If

participants' beliefs were shaped only by the content of the forecast, and not by its format, then they should have reacted the same way to the forecast information regardless of its format. We expected, however, that forecast formats would matter to the participants, resulting in different attitudinal reactions to presumably equivalent information.

5.1. Method

Participants (N = 225 valid participants from Amazon Mechanical Turk, or AMT, after excluding those with duplicate IP addresses and failed attention checks, which was done in all the studies) saw an election forecast that displayed either the chance of Clinton winning (i.e., the chance-forecast-displayed condition) or Clinton's predicted margin of victory (i.e., the margin-forecast-displayed condition). This study was conducted four days before the election, when Clinton led Trump 51.5% vs. 48.5%, and Clinton was projected to have a 64% chance of winning, according to *fivethirtyeight.com*.

Participants were asked to rate their opinion of the prediction on a scale from 1 (“*very good news*”) to 7 (“*very bad news*”), along with answering several other questions (see Appendix 1). Since a wider perceived lead for Clinton should be interpreted as better news by her supporters but worse news by Trump supporters, we coded the extremity of each participant's attitude as the absolute difference from the midpoint of the scale (extremity = $|\text{rating} - 4|$).

5.2. Results and Discussion

Participants in the chance-forecast-displayed condition had more extreme reactions to the state of the election conveyed by the forecast than participants in the margin-forecast-displayed condition ($M_{\text{chance-forecast-displayed}} = 1.76$ vs. $M_{\text{margin-forecast-displayed}} = 1.38$, $d = 0.32$, $t(223) = 2.41$, $p =$

.017). Since the forecasts were presumably statistically equivalent, the diverging attitudes were attributable to the difference in forecast format. In particular, these results suggest that chance forecasts are seen as conveying a stronger lead than margin forecasts convey. We replicated this finding in an additional study (Study A1 in Appendix 1) in which people were presented with information about changes in forecasts (rather than static forecasts) in both formats.

Notably, this effect on attitudes cannot be explained by anchoring-and-adjustment, which postulates that participants start their estimation process with a target value and then adjust the value upward or downward (Tversky & Kahneman, 1974). Because the forecast values were 51.5 in the chance-forecast-displayed condition and 64 in the margin-forecast-displayed condition, it was impossible for participants to adopt either value as a starting target value to generate a response on a 1–7 scale. Furthermore, since the forecasts and participants' responses used different units, neither established theories nor existing empirics about anchoring would predict our findings (Mochon & Frederick, 2013). Nonetheless, to be thorough, we test for use of anchoring as a strategy in Studies 5 and 6, using self-report data.

In sum, Study 1 shows, based on the same attitudinal outcome measurement in both forecast conditions, that the extremity of attitudes varies with forecast format. However, because Study 1 used a subjective outcome measure, the results provide only an indirect test of how people differentially understand the forecasts. In particular, Study 1 does not identify the source of the difference in attitudes or establish whether participants develop better-calibrated opinions with one forecast format over the other.

As a more direct approach, in the remaining studies in the article (except for Study 4), we tested for a specific form of forecast-format bias, a *chance-margin discrepancy* between participants' understanding of margins from chance forecasts and their understanding of chance

from margin forecasts. In this approach, we either displayed the chance forecast and asked participants to estimate the corresponding margin forecast, or displayed the margin forecast and asked participants to estimate the corresponding chance forecast. This approach enabled us to determine the accuracy of participants' estimates separately from each forecast format..

6. Study 2: Presidential Election Forecasts of Winning Chance vs. Vote

Margin

Study 1 demonstrates that presumably equivalent forecasts communicated in different formats yield different attitudes about the election. We contend that this occurs because untrained people do not interpret the forecasts correctly. Instead, a systematic bias in estimation occurs due to a lack of understanding about the non-linear relationship between the forecast formats (Figure 1). Based on this lack of understanding, we propose that people perceive chance forecasts as more extreme than equivalent margin forecasts from the same predicted distribution. As a result, we predict that people will overestimate margins when shown chance forecasts and underestimate chances when shown margin forecasts.

6.1. Method

We examined the chance-margin discrepancy over the course of the 2016 U.S. presidential election by presenting voting-age participants with either a chance forecast (probability of a specific election outcome) or a presumably equivalent margin forecast (predicted vote share of one candidate) and asking participants to estimate the forecast in the other format. If people's beliefs are shaped only by the content of a forecast, regardless of its

format, then they should be able to use one forecast to estimate the equivalent forecast in the alternative format, without systematic errors.

We collected data in six waves between August and November 2016, prior to the 2016 election. We collected completed surveys from a total of 1,163 U.S. residents, recruited from AMT. As in Study 1, we used the latest election forecasts posted on the website *fivethirtyeight.com* at the time of each wave. We showed participants one of two forecast formats: (1) the chance (probability) of each candidate winning or (2) the predicted margin (difference in vote shares) by which the favored candidate beats the other. Both forecasts were generated from the same statistical model, based on the same aggregated polling data, and therefore summarized the same predicted distribution of outcomes. The results of the election polls used by *fivethirtyeight.com* to generate the predicted distribution varied over time, yielding different forecasts used in each of the six waves.

Participants in each wave were randomly assigned to one of two conditions. In the chance-forecast-displayed condition, participants saw the probability of each candidate winning and were asked to estimate the corresponding margin forecast. In the margin-forecast-displayed condition, participants saw the predicted vote shares for the two candidates and were asked to estimate the corresponding chance forecast for each candidate. We counterbalanced the order in which the candidates were presented in all waves except the first, and we confirmed that the order of presentation did not affect the results.

6.2. Results

Participants' estimates were significantly biased by forecast format (Figure 2), suggesting that they did not interpret the two forecasts formats as representing the same underlying political

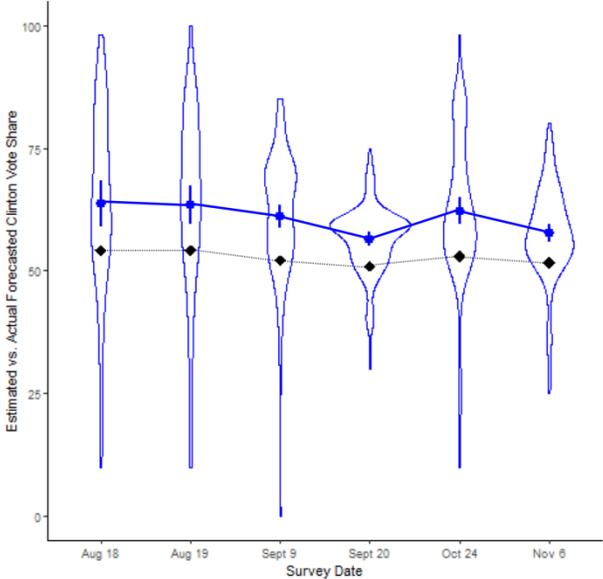
situation. Overall, participants in the chance-forecast-displayed condition (displaying an average 74.2% chance of Clinton winning) overestimated the margin forecast (60.5% estimated vote share for Clinton vs. 52.6% actual forecasted vote share, $d = 0.58$, $t(575) = 13.91$, $p < .001$). Participants in the margin-forecast-displayed condition (displaying an average vote share of 52.6% for Clinton) underestimated the chance forecast (59.5% estimated chance of Clinton winning vs. 74.2% actual forecasted chance, $d = 0.87$, $t(586) = 21.16$, $p < .001$).

Both forecast formats yielded significantly biased estimates, but the estimates were biased in opposite directions—while participants underestimated the chance forecasts, they overestimated the margin forecasts. These findings replicated as the actual forecasts varied across each of the six waves (all $ps < .001$), establishing the robustness of the chance-margin discrepancy. Furthermore, the vast majority of participants displayed the chance-margin discrepancy—80% of participants in the chance-forecast-displayed condition overestimated the margin forecast, while 82% of participants in the margin-forecast-displayed condition underestimated the chance forecast. This was not explained by participants simply reporting back the provided value, as the findings persisted when excluding such estimates.

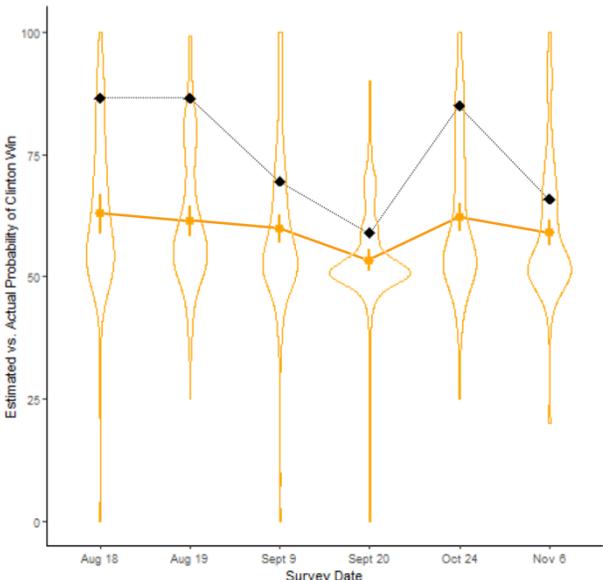
Figure 2: Estimation results based on different forecast formats in Study 2

Colored dots indicate participants' mean estimates (margin estimates in blue; chance estimates in orange), made after viewing the other forecast format; error bars represent the confidence intervals, and the colored plots show the distribution of estimates. By contrast, the black diamonds represent the actual election forecast at the time of each survey.

Chance Forecast Displayed



Margin Forecast Displayed



Despite the strong chance-margin discrepancy, participants in both conditions were sensitive to the truth of the forecast. A regression predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value ($\beta_{\text{ACTUAL}} = 2.18$, $\eta^2 = 0.032$, $t(574) = 4.35$; $p < .001$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value ($\beta_{\text{ACTUAL}} = 0.26$, $\eta^2 = 0.034$, $t(585) = 4.54$; $p < .001$). Furthermore, the magnitude of the chance-margin discrepancy depended on the actual forecast value: we found a significant interaction between estimation condition and actual chance forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.88$, partial $\eta^2 = 0.099$, $t(1159) = 11.34$; $p < .001$) in a follow-up regression predicting estimation error.

These results suggest that while participants' estimates were sensitive to changes in the state of the election over time, the robust chance-margin discrepancy (greater underestimation in the margin-forecast-displayed condition and greater overestimation in the chance-forecast-displayed condition) was larger when the actual forecast was more extreme. We replicated these findings in a similar study that instead used state-level chance and margin presidential election forecasts for four non-identified states (see Appendix 2).

6.3. Discussion

Our findings about the impact of forecast format on people's interpretation of election forecasts rely on the assumption that the two summary statistics (chance and margin) are in fact equivalent, meaning that and the statistics represent the same underlying prediction distribution and the assumed shape of the distribution is known and accurate. During the election, however, forecasters disagreed with each other about how to translate each candidate's polling-based,

forecasted vote share (margin) into their chance of winning the election. Essentially, forecaster disagreement about the likelihood of substantial changes in voter preference occurring in the time remaining before the election yielded different estimates of the tails of the distribution for the same forecasted margin. Furthermore, the distribution of votes across states and the role of the Electoral College complicated chance predictions.

We selected *fivethirtyeight.com* because their chance predictions were the most conservative (e.g., compared to the Princeton Election Consortium) and their forecasts have, in fact, been unbiased overall (Silver, 2019). However, the 2016 U.S. presidential election did not turn out as predicted, and evidence suggests that polling aggregation may not have been optimal, overweighting past polls and insufficiently updating over time, based on changes in voter preferences (Wright & Wright, 2018). Thus, we cannot exclude the possibility—however unlikely—that our findings are explained by the superior accuracy of lay people’s estimates over the professional forecasts for this specific election. Next, we extended our investigation into another forecast domain that does not share the limitations of this specific election and enables us to assess the generality of the chance-margin discrepancy.

7. Study 3: Sports Forecasts of Winning Chances vs. Point-Spreads

In this study, we tested the effect of forecast format in a different, pervasive, and important forecast domain—sports forecasts. By using this domain, we can generalize our findings to a setting in which chance and margin forecasts are quantified on different scales.

7.1. Method

In the pre-registered Study 3 (AsPredicted #6391, N = 258 valid AMT participants), we compared inferences about NBA basketball games. Specifically, participants viewed either the predicted chances of winning or the predicted margin of winning (point-spread) and then estimated the other forecast format, as in Study 2. We used chance and point-spread forecasts from *fivethirtyeight.com* for four basketball games that took place in October 2017, and we collected data shortly before the games were played. This kind of sports forecast has a high demonstrated accuracy (Spann & Skiera, 2009).

The matches ranged from almost even, with neither team strongly favored (Rockets vs. Hornets: 54% vs. 46%, point spread of 1), to very uneven, with one team highly favored (Warriors vs. Wizards: 84% vs. 16%, point spread of 10). The two other games lay in between (Nuggets vs. Hawks: 70% vs. 30%, 5.5 points; Knicks vs. Nets: 59% vs. 41%, 2.5 points). Unlike the election forecasts used in Study 2, all four forecasts correctly predicted the subsequent game outcomes.¹ Participants in the chance-forecast-displayed condition saw the forecasted chance that each team would win for each of the four games and then estimated the corresponding forecasted point-spreads; participants in the margin-forecast-displayed condition saw the point spread for each game and then estimated the corresponding forecasted chances of winning.

¹ Retrieved from https://www.basketball-reference.com/leagues/NBA_2018_games.html

An additional advantage of this context is that point spread and chance are measured on different scales (number of points and percentage, respectively). We counterbalanced the order in which the teams were presented and also varied the framing (e.g., Warriors winning vs. Wizards losing) and confirmed that neither team order nor outcome framing affected the results. As in Study 2, participants were instructed to estimate what the website predicted rather than to state their own beliefs.

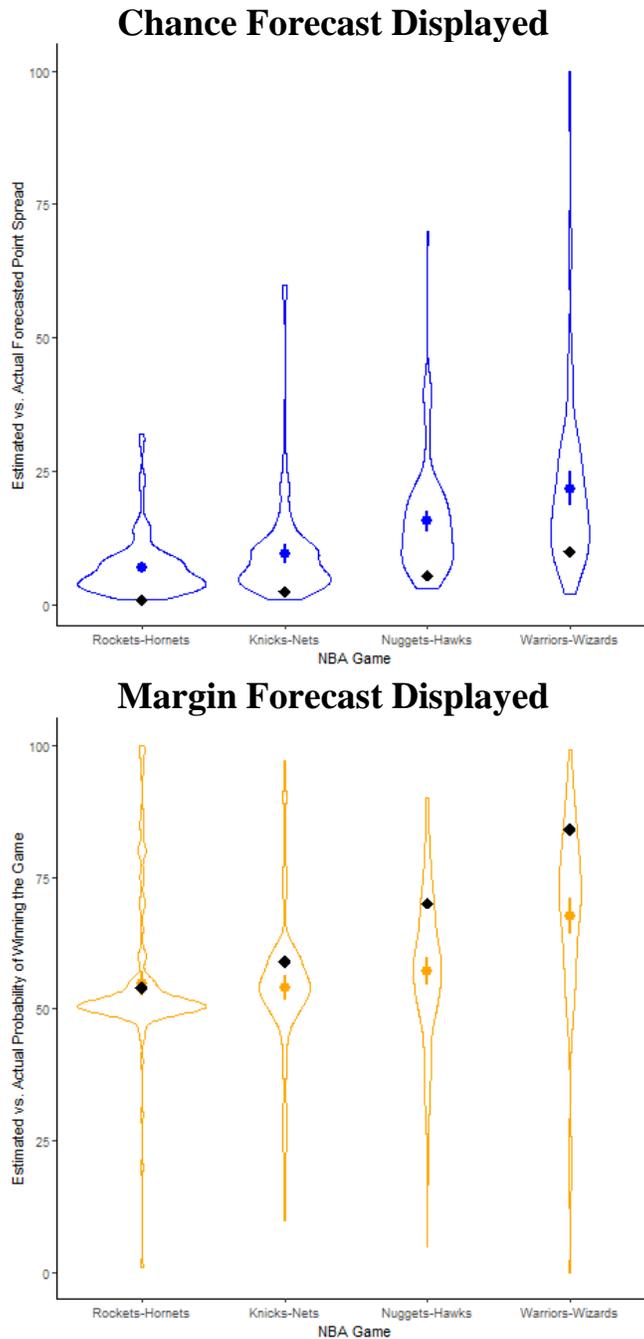
7.2. Results and Discussion

Figure 3 presents the participants' estimation results. Participants in the chance-forecast-displayed condition overestimated the point spread for all four games (all p s < .001), with an average overestimation of +8.73 (13.48 estimated vs. 4.75 actual, $d = 0.94$, $t(102) = 9.49$, $p < .001$). By contrast, participants in the margin-forecast-displayed condition significantly underestimated the chance that the favored team would win in every game (p s < .001) except for in the Rockets-Hornets game ($p = .517$), with an average underestimation of -8.35 percentage points (58.40 estimated vs. 66.75 actual, $d = 1.11$, $t(117) = 12.11$, $p < .001$).

This chance-margin discrepancy was pervasive across participants—93% of participants in the chance-forecast-displayed condition overestimated the point-spread, and 88% of participants in the margin-forecast-displayed condition underestimated the chance that the favored team would win. This was not explained by participants simply reporting back the provided value, as the findings persisted when excluding such estimates. The chance-margin discrepancy also was not explained by lack of familiarity with the context, as self-reported level of involvement with NBA games did not moderate the estimation errors.

Figure 3: Estimation results based on different forecast formats in Study 3

Colored dots indicate participants' mean estimates (margin estimates in blue; chance estimates in orange), made after viewing the other forecast format; error bars represent the confidence intervals, and the colored plots show the distribution of estimates. By contrast, the black diamonds represent the actual forecasts for each NBA game.



Participants did not ignore the forecasts in making their estimates. A regression with clustered standard errors predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value ($\beta_{\text{ACTUAL}} = 1.66$, $\eta^2 = 0.21$, $t(410) = 11.51$, $p < .001$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value ($\beta_{\text{ACTUAL}} = 0.44$, $\eta^2 = 0.10$, $t(470) = 7.15$, $p < .001$). We again found a significant interaction between estimation condition and actual chance forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.76$, partial $\eta^2 = 0.097$, $t(880) = 10.12$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that while participants' estimates were sensitive to the differences in the provided forecasts, participants nevertheless displayed a robust chance-margin discrepancy that was more pronounced with more extreme forecasts.

These results extend our findings of the chance-margin discrepancy from political election forecasts to sports game forecasts. Participants in the chance-forecast-displayed condition overestimated point-spreads, while participants in the margin-forecast-displayed condition (i.e. those shown point-spread forecasts) underestimated chances of winning. We found similar but somewhat weaker results in two additional studies in which participants made only a single forecast estimate, this time about a football game (see Appendix 3).

8. Study 4: Impact of Forecast Format on Risk Taking

Study 3 demonstrated the effect of forecast format on beliefs about the outcomes of NBA games. Could such forecast-induced beliefs impact consequential risk-taking decisions? Study 4

tests whether forecast format impacts people's willingness to bet on the team predicted to win the upcoming game.

8.1. Method

In the pre-registered Study 4 (AsPredicted #8906, $N = 371$ valid AMT participants), participants read about an upcoming NBA game in which one team was strongly favored to win (Toronto Raptors with 93% chance vs. Atlanta Hawks with 7% chance; Toronto Raptors to win with a point spread of 16). Participants saw the team names and either the forecasted chance that each team would win (in the chance-forecast-displayed condition) or the forecasted point spread by which the Raptors would win (in the margin-forecast-displayed condition).

Participants were told that they had been entered into a lottery, and five winners would receive a \$5 bonus each. Before finding out if they had won the lottery, participants could choose to bet as much of the potential \$5 as they wished on the team predicted to win (Raptors) and keep the remainder. For each \$1 bet on the winning team, participants would receive \$3; participants would lose all money bet on the losing team. Participants then indicated how much of the \$5 they wished to bet. After the study was completed, the lottery winners were notified and paid based on their bets.

8.2. Results

Participants bet more in the chance-forecast-displayed condition than in the margin-forecast-displayed condition ($M_{\text{chance-forecast-displayed}} = \3.03 vs. $M_{\text{margin-forecast-displayed}} = \2.62 , $d = .22$, $t(369) = 2.15$, $p = .033$). While there was no significant difference in the likelihood of betting any money (80% vs. 84%, $\chi^2 = .78$, Fisher $p = .35$), there was a significant difference in

the amount wagered among the 304 participants who *did* bet on the Raptors (rather than keep all their money)—participants who saw the chance forecast wagered more money than participants who saw the margin forecast ($M_{\text{chance-forecast-displayed}} = \3.61 vs. $M_{\text{margin-forecast-displayed}} = \3.28 , $d = 0.23$, $t(302) = 2.01$, $p = .045$). The effect of forecast format on betting did not differ across different levels of basketball experience (e.g., knowledge about and interest in NBA basketball, number of games watched). This suggests that the impact of forecast format on betting was not due to a lack of relevant knowledge. Rather, the chance (vs. margin) forecast led to higher confidence about the outcome, as expressed by participants' willingness to bet more on the game. We find similar results across seven other studies (total $N = 1,999$) summarized in Appendix 4.

9. Study 5: Accuracy of Estimates Under Epistemic Uncertainty

The studies thus far suggest that people treat presumably equivalent forecasts differently, as if chance forecasts (both for an upcoming election and for sports outcomes) convey a more polarized reality than margin forecasts. However, the conclusions from the aforementioned studies all rest on the assumption that the chance and margin forecasts are generated from a well-calibrated full outcome distribution, so that the forecasts are in fact equivalent. The contexts tested thus far involve aleatory uncertainty under which the ground truth is not knowable at the time the forecast is made. In some sense, aleatory uncertainty makes those studies a conservative test of our hypotheses, as subjective probability estimates under aleatory uncertainty are typically less extreme than those under the more knowable epistemic uncertainty (Tannenbaum, Fox, & Ulkumen, 2016).

One consequence of using domains with aleatory uncertainty is that we can never know whether the forecasts were correct, particularly given there is always uncertainty about the assumptions underlying the forecasts. This is highlighted in the election context—we cannot know whether Donald Trump’s victory in the 2016 U.S. presidential election was evidence that the forecasts were mis-calibrated, or whether the forecasts were indeed accurate and Trump’s victory simply reflected the nonzero probability that he would win (e.g., Wright & Wright, 2018). While *fivethirtyeight.com* forecasts have been shown to be well-calibrated overall (Silver, 2019), any specific forecast selected to be used in a study could have been mis-calibrated.

To address this issue, in this study, we instead test the effect of margin vs. chance forecasts in a context (movie ratings by genre) involving epistemic uncertainty—in other words, we construct forecasts about questions for which an objective ground truth (already published movie ratings) exists. The fact that this ground truth is available enables us to construct a representative sample of statistically “correct” forecasts. Thus, we extend our findings to situations in which people do have the opportunity, at least in principle, to be exposed to the information that would enable them to generate accurate prior beliefs (i.e., beliefs before seeing the forecast).

9.1. Method

In the pre-registered Study 5 (As Predicted #33246, N = 400 valid AMT participants), participants completed 10 tasks in which they saw either a series of chance forecasts or a series of margin forecasts involving a comparison between ratings of movies in different genres, and participants estimated the corresponding value for the other forecast format. To generate the forecasts, we collected ratings from *Metacritic.com* for 7,746 movies in five genres (action,

adventure, biography, crime, and horror). Metacritic standardizes and aggregates the assessments of movie critics on a 0 to 100 scale for each movie. This is a widely-used resource for entertainment ratings; during the time period when the study was run, Metacritic received between 15 and 18 million visits per month.² Since our data effectively constitutes a census with many observations (over 1000) in each genre, the Metacritic data provides the ground truth for comparisons of movie ratings by genre.

Stimuli. We generated our own forecasts for use in the study (unlike in the prior studies, which used publicly available forecasts generated by others). Our goal was to create a well-calibrated chance forecast (the probability that movies in one genre are, on average, rated higher than movies in another genre) and an equivalent margin forecast (the expected difference in the average movie ratings between the same two genres). For each of the 10 pairs of genres, we generated 200 random samples (without replacement) of 30 movies (15 in each genre). Each of the samples formed the basis of a pair of equivalent chance and margin forecasts. We generated the forecasts by conducting a Bayesian linear regression (see Argyropoulos, 2013, and the code in Appendix 5) to predict movie ratings based on genre for each sample, with uniform priors on the coefficient. The Bayesian linear regression then yielded an estimated posterior distribution (i.e., forecasted distribution based on the data in the sample) of the difference in ratings between the two genres.

We treated this posterior distribution as a representation of the beliefs of an optimal Bayesian forecaster with no prior beliefs about differences between genres, who observes a

² <https://www.similarweb.com/website/metacritic.com>; Retrieved on 11/27/2019.

specific sample of movies from the two genres in question. We then generated optimal Bayesian forecasts, conditional on the observed sample, of (a) the chance that one genre is higher rated than the other (by computing the corresponding tail of the posterior distribution) and (b) the expected margin by which one genre's average rating exceeds the other (as the median of the posterior distribution of genre differences).

Procedures. In the study, participants were randomly assigned to either the chance-forecast-displayed or margin-forecast-displayed condition. Participants were also randomly assigned to be shown optimal Bayesian forecasts we had generated, for either unidentified comparisons (e.g., Genre A vs. Genre B) or identified comparisons (e.g., action vs. adventure). As we discuss subsequently, this enabled us to test hypotheses about the participants' use of priors.

Participants were provided with contextual information about the movie genre ratings to enable comprehension of the forecasts. In addition, participants were trained to interpret these forecasts, which was particularly important since they would not have been exposed to movie genre forecasts previously (unlike election and sports forecasts). The survey began with a series of instructional tasks. Participants were told about the range (0–100), average (54) and standard deviation (18) of the movie ratings, with the standard deviation explained as the range that encompassed approximately 66% of the values. Participants were required to report this information back accurately before proceeding.

Before the study, we tutored and trained the participants about the forecasting task. Participants first learned about how the forecasts were constructed, including the relationship between chance and margin forecasts: that a 50% chance forecast corresponded to a margin forecast of zero, and that a chance forecast above 50% corresponded to a forecasted margin

above zero. Then, they had to answer four interpretation questions correctly before proceeding. Lastly, participants completed a practice task, formatted like the actual tasks, that involving translating a 50% chance forecast to a margin forecast of zero, or vice versa, depending on the condition. Participants who gave an incorrect answer were given feedback and an opportunity to correct their answer to the practice task.

After completing the instructional section, each participant completed 10 tasks involving all possible pairs of genres (in random order). In each task, one of the 200 pairs of chance and margin forecasts for that genre pair (each computed based on a unique random sample of 30 movies) was selected at random. Participants either were shown the chance forecast (the probability that one genre had a higher average rating than the other) and estimated the corresponding margin forecast (the expected difference in average ratings between the two genres), or vice versa, depending on condition.

In each task, participants whose estimate was in the opposite direction of the forecast (e.g., a participant viewed a margin forecast in which Genre A was rated higher than Genre B, but estimated a chance $< 50\%$ that Genre A was rated higher than Genre B) were then asked a follow-up question in which the discrepancy was explained, giving participants the option (but not a requirement) to update their forecast. All participants were informed that one of their 10 tasks would be selected at random and evaluated for accuracy, yielding a bonus of up to \$1 with a linear incentive for estimates within 25 of the actual forecast value.

Methodological Significance. While much more complicated than in the prior studies, the approach we adopted in Study 5 holds several important advantages. First, the information presented to participants constitutes a representative set of sampling-based forecasts, and therefore guards against the possibility that our effects are overestimated by stimuli selection.

Second, instead of relying on forecasts from inaccessible generation methods, we generated the forecasts ourselves using a Bayesian model that provides full transparency as to the methods. Third, the forecasts presented to participants were based on different random samples from a known population of movies. This allowed us to directly test the possibility that participants provided non-equivalent forecasts because they had additional information that enabled them to make *more accurate* forecasts. Fourth, by experimentally manipulating participants' knowledge of the comparison at hand (i.e., whether they knew the identity of the genres), we had the opportunity to examine the impact of participants' ability to use their prior beliefs. Lastly, given that forecasts were generated from a known population, we were able to provide participants with more information about the underlying distribution, including the standard deviation of the ratings in question.

9.2. Results and Discussion

Out of a total of 400 valid participants, 220 gave correct estimates in the practice task (equating a 50% chance to a zero-difference margin). The low success rate reflects the difficulty of this task, particularly in the context of comparing movie genre ratings—in which participants were unlikely to have previously encountered either margin or chance forecasts. As a conservative test, we report results including only these 220 participants, thus testing the interpretation of chance and margin forecasts among individuals with the highest demonstrated baseline competence. The results using all participants are provided in Appendix 6 and reveal an even greater chance-margin discrepancy.

Chance-Margin Discrepancy. Table 1 shows estimation results for each pair of genres by condition. Participants in the chance-forecast-displayed conditions consistently overestimated

the margin prediction in each pair of genres (all $ps < .001$) by an average of 19.2 rating points overall ($M_{\text{estimated}} = 28.1$ vs. $M_{\text{actual}} = 8.9$, $d = 0.94$, $t(789) = 21.7$, $p < .001$). Conversely, participants in the margin-forecast-displayed conditions consistently underestimated the chance prediction in each pair of genres (all $ps < .001$) by an average of 21.1 percentage points overall ($M_{\text{estimated}} = 62.2$ vs. $M_{\text{actual}} = 83.3$, $d = 1.29$, $t(1409) = 44.7$, $p < .001$). Moreover, a majority of participants (88%) in the chance-forecast-displayed condition overestimated the expected difference in ratings (i.e., in Panel 1 of Figure 4, most data points are above the dashed line, which represents the actual corresponding margin forecasts). Likewise, a majority of participants (91%) in the margin-forecast-displayed condition underestimated the chance that the favored genre would be higher rated (i.e., in Panel 2 of Figure 4, most data points are below the dashed line, which represents the actual corresponding chance forecasts). As in the prior studies, the chance-margin discrepancy occurred because both forecast formats yielded biased estimates, but the estimates were biased in opposite directions.

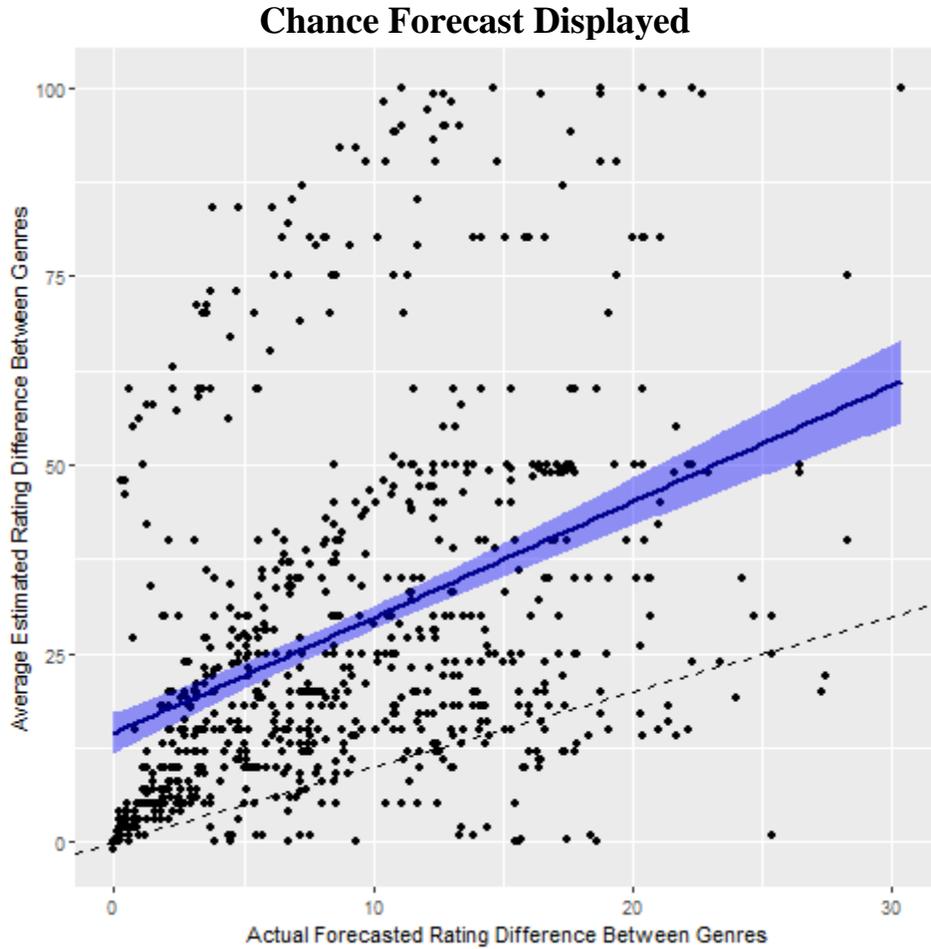
Table 1: Estimation errors for each pair of genres in Study 5

Genres	Chance Forecast Displayed			Margin Forecast Displayed		
	Est. Margin	Avg. Margin	Error	Est. Chance	Avg. Chance	Error
Action vs. horror	20.1	4.3	+15.8***	55.9	71.5	-15.5***
Adventure vs. crime	22.8	5.7	+17.1***	59.3	75.5	-16.2***
Action vs. adventure	25.7	6.2	+19.5***	58.8	78.2	-19.3***
Action vs. crime	23.3	6.5	+16.8***	61.5	82.3	-20.7***
Adventure vs. horror	27.7	7.9	+19.8***	62.1	82.1	-19.9***
Crime vs. horror	25.1	8.2	+16.9***	61.6	79.6	-18.0***
Biography vs. crime	30.0	9.3	+20.7***	62.3	86.0	-23.7***
Adventure vs. biography	32.8	10.6	+22.2***	61.9	86.9	-25.0***
Action vs. biography	36.5	14.0	+22.5***	68.4	94.1	-25.8***
Biography vs. horror	37.2	16.2	+20.9***	70.4	97.1	-26.7***

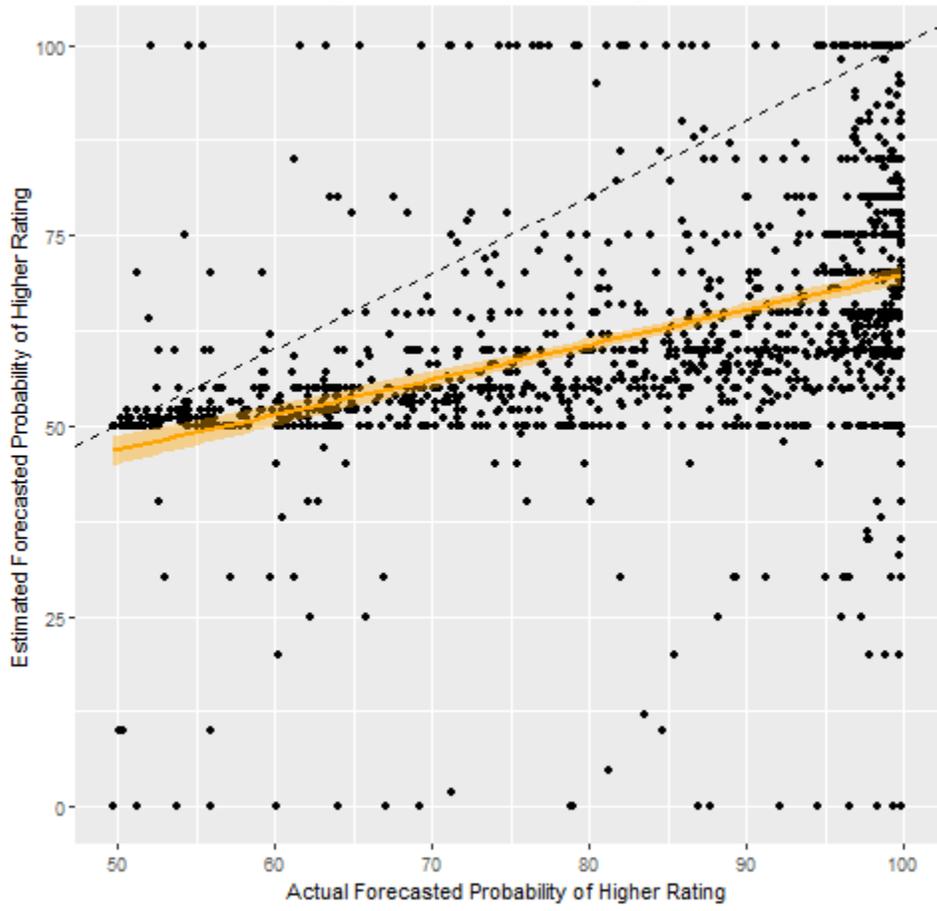
Notes: *** indicates the error is significantly different from zero, $p < .001$.

Figure 4: Estimation results based on different forecast formats in Study 5

The colored lines indicate participants' mean estimates (margin estimates in blue; chance estimates in orange), made after viewing the other forecast format; the confidence intervals are shaded. By contrast, the dashed black lines represent the actual forecasts we generated from the random samples drawn from all extant ratings. The estimates of each individual participant appear as black diamonds.



Margin Forecast Displayed



As shown in Figure 4, participants' degree of underestimation of chance and overestimation of margin increased with the extremity of the provided forecast. A regression (with clustered standard errors at the person level) predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value ($\beta_{\text{ACTUAL}} = 1.53$, $\eta^2 = 0.159$, $t(788) = 9.39$, $p = .013$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value ($\beta_{\text{ACTUAL}} = 0.46$, $\eta^2 = 0.172$, $t(1408) = 14.0$, $p < .001$). We found a significant interaction between estimation condition and actual margin forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -1.56$, partial $\eta^2 = 0.057$, $t(2196) = 8.4$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that while participants' estimates were sensitive to the differences in the provided forecasts across sample results, participants nevertheless displayed a robust forecast-format bias that was more pronounced with more extreme forecasts.

Estimation Process. In the case of elections and NBA games, it is possible that some participants under-relied on the provided forecasts—and hence gave inconsistent estimates—because they factored in their own knowledge about the presidential candidates or NBA teams, thus potentially improving the accuracy of the forecast. (In Bayesian terms, these participants used their own prior beliefs rather than the prior of the forecaster.) In this study, we assumed a uniform prior distribution on the coefficient representing the difference in ratings between a pair of genres. However, participants might still have their own prior beliefs (distinct from the uniform prior) on the difference in ratings between a pair of genres, and—as in the earlier studies—this could contribute to the chance-margin discrepancy.

Two methodological features of this study were designed to address this possibility. First, we experimentally manipulated whether or not participants were informed about the identities of the genres in question, thereby enabling or preventing the use of their own idiosyncratic prior beliefs. If participants failed to estimate the forecasts because of a (potentially normative) reliance on their own priors, then the chance-margin discrepancy and the estimation error should be reduced among participants who did not know the genre identities.

Second, because we randomly assigned each participant to one of many pairs of forecasts (i.e., corresponding to different samples of movie ratings observed by the forecaster), we can distinguish between the forecasted margin and the true difference. Thus, we can compare participants' average forecast estimate to the true difference in ratings for that genre-pair in the full Metacritic data. If participants relied on their own well-calibrated prior beliefs about the genres, then participants who were provided with the genre identities should have made forecast estimates that were biased specifically in the direction of the true value.

Overall, we found little evidence that the chance-margin discrepancy can be explained by participants' reliance on their prior beliefs. The accuracy of participants' estimates, relative to the corresponding forecasts, did not vary significantly between participants who saw specific genre information (e.g., action vs. adventure, facilitating the use of idiosyncratic priors) and those who did not (e.g., Genre A vs. Genre B). Participants in the chance-forecast-displayed conditions overestimated the margin forecast to a similar degree regardless of whether they saw the genre identities ($\text{Error}_{\text{Genre}} = 20.9$ vs. $\text{Error}_{\text{No-Genre}} = 17.7$, $d = .145$, $t(788) = 0.8$, $p = .429$). Likewise, participants in the margin-forecast-displayed conditions underestimated the chance forecast to a similar degree in both conditions ($\text{Error}_{\text{Genre}} = -20.7$ vs. $\text{Error}_{\text{No-Genre}} = -21.5$, $d = .042$, $t(1408) = 0.4$, $p = .711$).

Next, we tested whether participants in the genre-information conditions based their answers on not only the provided forecasted differences in ratings but also the (unobserved) true differences in genre ratings. Participants' margin forecast estimates were significantly predicted by the actual, sample-based, forecasted margins ($\beta_{\text{ACTUAL}} = 1.43$, $t(377) = 5.93$, $p < .001$) but not by the true difference for the genre pair in the full data ($\beta_{\text{TRUE}} = 0.63$, $t(377) = 0.78$, $p = .436$). Likewise, participants' chance forecast estimates were significantly predicted by the actual, sample-based, forecasted margins ($\beta_{\text{ACTUAL}} = 1.21$, $t(737) = 11.47$, $p < .001$) but not by the true difference for the genre pair in the full data ($\beta_{\text{TRUE}} = .63$, $t(737) = 1.22$, $p = .224$). Thus, the observed misestimation of forecasts cannot be explained by participants with well-calibrated priors adjusting certain provided forecasts (i.e., those derived from non-representative samples) to be more accurate to the true differences in movie ratings across genres.

Thus far, we have discussed potential theoretical accounts for the chance-margin discrepancy, particularly normative accounts that might describe participants' estimation, and we found little support in the data. Next, we turn to a descriptive question—what strategies are participants using? We consider whether some strategies are associated with less bias. To this end, we analyzed participants' open-ended descriptions of the strategies they used, after coding their responses into categories (Table 2).

Table 2: Participants' self-reported decision strategies in Study 5

	All	Chance Forecast Displayed	Margin Forecast Displayed
Nonspecific intuition	46.3%	48.1%	45.4%
Anchoring and adjusting	11.8%	11.4%	12.1%
Using standard deviation or statistical distribution	10.9%	19.0%	6.4%
Using mathematics and the forecast	9.1%	6.3%	10.6%
Taking a random guess	4.5%	3.8%	5.0%
Assuming chances and share were the same	1.8%	2.5%	1.4%
Ranking the possible outcomes	0.5%	0%	0.7%
Estimating near the average	0.5%	0%	0.7%
Using own opinion of genres	0.5%	0%	0.7%
Other	0.9%	2.5%	0%
No strategy provided	13.2%	6.3%	17.0%

Overall, 71.4% of participants indicated in their response that they had used the forecast provided, and this did not significantly differ between the chance-forecast-displayed and margin-forecast-displayed conditions. However, the ways in which participants used the forecasts varied widely, as did the level of detail they provided about their approach to making their estimate. Only 38% of participants indicated that their strategy involved scaling higher forecasts into higher estimates, with no significant difference between the chance-forecast-displayed and margin-forecast-displayed conditions.

The specific strategies, as shown in Table 2, weakly differed across conditions ($\chi^2 = 18.6$, $df = 10$, $p = .045$). Relatively few participants described completely ineffective strategies, such as random guessing (4.5%) or adopting the provided forecast as the estimate (1.8%). The most common answer (46.3%) described a non-specific, intuition-based strategy that typically involved the provided forecast. Many participants described some kind of rough algorithm they had used to convert the provided forecast into the desired forecast. Some described anchoring and adjustment approaches (11.8%), such as adding 50 to the margin forecast to generate a chance forecast. Some described other ad hoc mathematical transformations of the forecast (10.9%), such as doubling the margin forecast and adding it to 50 to get the chance forecast. Relatively few mentioned the standard deviation or discussed statistical distributions (10.9%), and those who did tended to describe intuitive approaches or self-described failed attempts to use statistical theory. Virtually no participants described anything recognizable as either Bayesian reasoning or even an attempt to map the types of forecasts to different summary statistics of an outcome distribution.

Overall, the results suggest that some strategies were less effective than others. Participants who mentioned using the forecast ($r = -.21$, $p = .002$) or scaling their estimates based

on the provided forecast ($r = -.14, p = .041$) had lower average absolute error. Furthermore, participants who mentioned using the standard deviation or a statistical distribution in some way had lower average absolute error ($r = -.23, p < .001$). That said, the systematic errors persisted regardless of the strategy used. In the chance-forecast-displayed condition, the margin forecast was overestimated by 18.8 percentage points among participants who mentioned using the forecast ($p < .001$) and by 8.1 percentage points among those who mentioned using the standard deviation or distribution ($p < .001$). Similarly, in the margin-forecast-displayed condition, the probability forecast was underestimated by 18.8 percentage points among participants who mentioned using the forecast ($p < .001$) and by 19.0 percentage points among those who mentioned using the standard deviation or distribution ($p < .001$).

Overall, the results of Study 5 confirm that the chance-margin discrepancy occurs even with a representative set of uniform-prior Bayesian forecasts, and the errors are not explained by participants' reliance on accurate prior beliefs. Even participants who reported using the seemingly-best strategies (among the commonly-reported strategies) overestimated the margins when shown chance forecasts and underestimated the chances when shown margin forecasts. We find similar (albeit noisier) results in a fully between-subjects replication study (Appendix 7).

10. Study 6: Accuracy of Estimates Based on Forecast Format in a Statistical Scenario

The findings of the studies thus far suggest that people treat equivalent forecasts differently, as if chance forecasts (for elections, sports and movie ratings) convey a more polarized reality than margin forecasts. However, concluding that the chance-margin discrepancy is a bias requires assuming that the chance and margin forecasts are generated from a full

outcome distribution that is not only correct but also at least theoretically knowable by the participants. In the election and sports contexts, this assumption may fail because the actual shape of the full distribution in reality may differ from the forecasters' assumptions, resulting in incompatible chance and margin forecasts. This was partially addressed in Study 5, in which the underlying truth (the distribution of all movie ratings) was known to the researcher, the forecasts were generated using uniform priors, and participants were provided with relevant descriptive information (e.g., the standard deviations of the ratings). Even in Study 5, however, participants might have lacked the information necessary to infer an accurate relationship between chance and margin forecasts. This leaves open the possibility that if participants have all the information necessary for an accurate, algorithmic calculation of one forecast from the other, the chance-margin discrepancy might be eliminated.

In Study 6, we addressed this issue by providing participants with sufficient objective information to enable a Bayesian to perfectly generate the forecast in one format from the forecast in the other format. We then tested the effects of forecast format on participants' judgments in an incentive-compatible statistical scenario. In addition, we measured participants' understanding of binomial distributions to test whether the results can be explained by systematically-biased beliefs about statistical distributions, and we asked participants to classify their estimation strategies.

10.1. Method

In the pre-registered Study 6 (AsPredicted #24681, $N = 578$ valid AMT participants), participants read a single statistical scenario involving a jar that contained an assortment of 99 marbles, some red and some green. The number of red marbles was chosen at random from a

uniform distribution. In the scenario, a sample (unobserved by the participant) was generated by drawing 20 marbles from the jar at random, with replacement, and recording the number of red marbles. According to the scenario, a statistician used the sample to determine both (1) the chance that there were more red than green marbles in the jar and (2) the expected margin (i.e., how many more red marbles than green marbles were in the jar).

To calculate the forecasts, we applied Bayes Theorem (see Appendix 8 for details). We can think of the margin forecast as the expected value of the posterior distribution of possible differences in the number of red and green marbles and the chance forecast as the mass of that posterior distribution above 50% red.

Participants were randomly assigned to see a forecast based one of two samples, containing either 12 or 14 red marbles out of 20, and expressed as one of the two forecast types. In the chance-forecast-displayed conditions, participants were shown one of two chance forecasts (81% or 96% chance that there were more red than green marbles in the jar) and estimated the corresponding margin forecast. In the margin-forecast-displayed condition, participants were shown one of two margin forecasts (18 or 36 more red than green marbles in the jar) and estimated the corresponding chance forecast. Each participant was paid a linear incentive for the accuracy of their estimate, with a maximum accuracy bonus of \$1 per person.

In addition, participants completed a probability distribution estimation task, adapted from Urminsky (2014). In the task, participants estimated the full binomial distribution arising specifically from the sampling process described in the scenario (20 samples with replacement from the jar). We counterbalanced the order of the binomial distribution estimation task and the forecast-based estimation to test whether prompting participants to consider the binomial

sampling distribution before making their forecast estimate would improve statistical thinking and reduce the forecast-format bias.

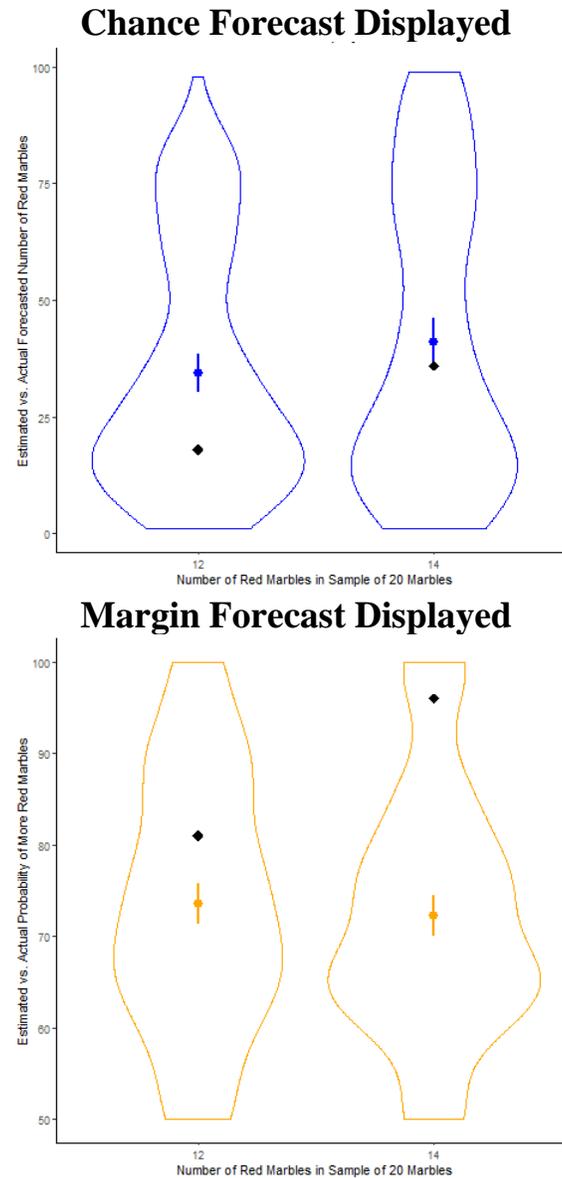
Lastly, after making their forecast-based estimate, participants were presented with verbal descriptions of possible decision strategies, similar to the types of strategies described in Study 5, and were asked to either choose the strategy that most closely aligned with their reasoning or else enter their own (open-ended) description. The descriptions provided to participants (see Appendix 9 for details) conveyed the following strategies: random guessing, using mathematics and the forecast, intuition, anchoring and adjusting either up or down, assuming the chance and the share were the same, and three approaches involving thinking in terms of statistical distributions: mentally simulating the sampling process, ranking the possible outcomes, and Bayesian reasoning.

10.2. Results and Discussion

As shown in Figure 5, participants in the chance-forecast-displayed conditions significantly overestimated the margin prediction when shown a forecasted chance of 81% ($p < .001$) but not when shown a forecasted chance of 96% ($p = .063$), with an average overestimation of 10.90 marbles ($M_{\text{estimated}} = 37.55$ vs. $M_{\text{actual}} = 26.65$, $d = 0.50$, $t(284) = 6.33$, $p < .001$). Conversely, participants in the margin-forecast-displayed condition significantly underestimated the chance that there were more red than green marbles in both scenarios ($ps < .001$), with an average underestimation of -15.51 percentage points ($M_{\text{estimated}} = 72.91$ vs. $M_{\text{actual}} = 88.42$, $d = 1.41$, $t(292) = 16.64$, $p < .001$). As in the prior studies, both forecast formats yielded biased estimates overall, but the estimates were biased in opposite directions.

Figure 5: Estimation results based on different forecast formats in Study 6

Colored dots indicate participants' mean estimates (margin estimates in blue; chance estimates in orange), made after viewing the other forecast format; error bars represent the confidence intervals, and the colored plots show the distribution of estimates. By contrast, the black diamonds represent the actual forecasts for each number of red marbles in the sample we generated from the population.



As in the prior studies, the chance-margin discrepancy was pervasive across participants, particularly in the margin-forecast-displayed conditions. Overall, 54% of participants in the chance-forecast-displayed conditions overestimated the margin of red vs. green marbles, while 81% of participants in the margin-forecast-displayed conditions underestimated the chance that there were more red than green marbles.

To test whether participants with more accurate statistical intuition in general are less prone to the forecast-format bias, we calculated the absolute error in the separate binomial distribution estimation task for each participant. We found no significant interaction between forecast format and absolute binomial-distribution error in predicting forecast estimation error ($\beta_{\text{FORMAT} \times \text{BIN_ERR}} = -0.07$, partial $\eta^2 = 0.001$, $t(574) = 0.74$, $p = .463$). We likewise found no significant interaction between forecast format and, specifically, overestimation of the tails of the binomial distribution in predicting forecast estimation error ($\beta_{\text{FORMAT} \times \text{TAILS}} = -0.07$, partial $\eta^2 = 0.002$, $t(574) = 1.06$, $p = .292$). These analyses further confirm that the observed bias is not explained by participants' generally mistaken beliefs about the shape of the binomial distribution.

Prompting participants to think about binomial distributions did not reliably improve forecast estimation. In the chance-forecast-displayed conditions, participants who completed the distribution task before the forecast estimation task had only slightly lower error (with vs. without distribution prompt: $M_{\text{error}} = 7.31$ vs. $M_{\text{error}} = 14.18$, $d = 0.24$, $t(283) = 2.00$, $p = .046$); there was no significant difference in the margin-forecast-displayed condition (with vs. without distribution prompt: $M_{\text{error}} = -16.19$ vs. $M_{\text{error}} = -14.81$, $d = 0.09$, $t(291) = 0.74$, $p = .459$).

Table 3: Participants' self-reported decision strategies in Study 6

	All	Chance Forecast Displayed	Margin Forecast Displayed
Nonspecific intuition	26.3%	23.5%	29.0%
Bayesian reasoning	14.0%	16.5%	11.6%
Using mathematics and the forecast	12.6%	13.0%	12.3%
Taking a random guess	11.9%	13.3%	10.6%
Assuming chance and share were the same	10.9%	10.9%	10.9%
Anchoring and adjusting (correct direction)	9.5%	8.1%	10.9%
Mentally simulating the sampling process	7.8%	9.5%	6.1%
Ranking the possible outcomes	3.1%	1.8%	4.4%
Anchoring and adjusting (incorrect direction)	0.9%	0.4%	1.4%
Other	2.9%	3.2%	2.7%

Lastly, we analyzed the descriptions participants selected for their estimation strategies (see Table 3 for the results and Appendix 9 for the text of the strategy options). Overall, the participants reported similar strategies across conditions (margin-forecast-displayed vs. chance-forecast-displayed, $\chi^2(9) = 13.54, p = .140$). A majority of participants did not report using a specific statistical strategy, instead using their intuition (26.3%), general mathematical reasoning (12.6%), or random guessing (11.9%). Among the statistical strategy descriptions, the description of Bayesian reasoning was selected by the most participants, but its absolute rate was still quite low (14.0%).

Overall, there was no significant relationship between the strategy used and the estimation error in either condition (chance-forecast-displayed: $F(1,283) = 2.20, p = .139$; margin-forecast-displayed: $F(1,291) = 0.15, p = .703$). In particular, participants who endorsed the Bayesian strategy had directionally but not significantly lower absolute error overall ($r = -.07, p = .118$), in both the chance-forecast-displayed condition ($r = -.09, p = .151$) and the margin-forecast-displayed condition ($r = -.07, p = .238$).

Participants who endorsed the intuition-based strategy had significantly lower absolute error overall ($r = -.11, p = .007$). This finding was driven by the margin-forecast-displayed condition ($r = -.14, p = .020$) but was not significant in the chance-forecast-displayed condition ($r = -.08, p = .151$). It may be that participants who reported using their intuition were more likely to focus on the forecast provided, as suggested by Study 5. However, we found no evidence that the chance-margin discrepancy was based on a flawed (and therefore potentially correctible) application of statistical reasoning principles, or that participants who attempted to use statistical reasoning did better. Instead, this study confirms the conclusions from Study 5: that people tend

to use ad-hoc heuristic reasoning, and that intuition-based approaches do not perform worse than statistics-based approaches.

10.3. Discussion

It is notable that we replicated the chance-margin discrepancy even though participants in this study were provided with full information and received an incentive for accuracy. As can be seen in the derivations in Appendix 8, applying Bayesian reasoning in this context is not computationally simple. However, if people truly are “intuitive Bayesians” (Griffiths & Tenenbaum, 2006; Lieder et al., 2012; Pouget et al., 2013), with mental processes that closely approximate optimal Bayesian reasoning, then we would have expected the systematic biases to be eliminated in this study, as we provided all necessary background information. The fact that the chance-margin discrepancy persisted in Study 6 suggests that the forecast-format bias lies in people’s mental processes and is not attributable to insufficient information as inputs to an optimal process. In fact, this finding is highly robust—we replicated the chance-margin discrepancy in two additional studies using the same full-information statistical scenario (Appendices 10 and 11).

11. General Discussion

Our research demonstrates that equivalent forecasts are consistently seen as more extreme when presented in terms of the chance of an outcome (e.g., the probability of a candidate winning an election) rather than in terms of the predicted margin (e.g., the expected difference in vote shares received by the winning vs. by the losing candidate). The findings, from six experiments reported here plus twelve replication experiments, are difficult to reconcile with

the view that people are skilled intuitive statisticians who are adept at Bayesian reasoning, but are consistent with prior findings that laypeople struggle to understand statistics and non-linear relationships.

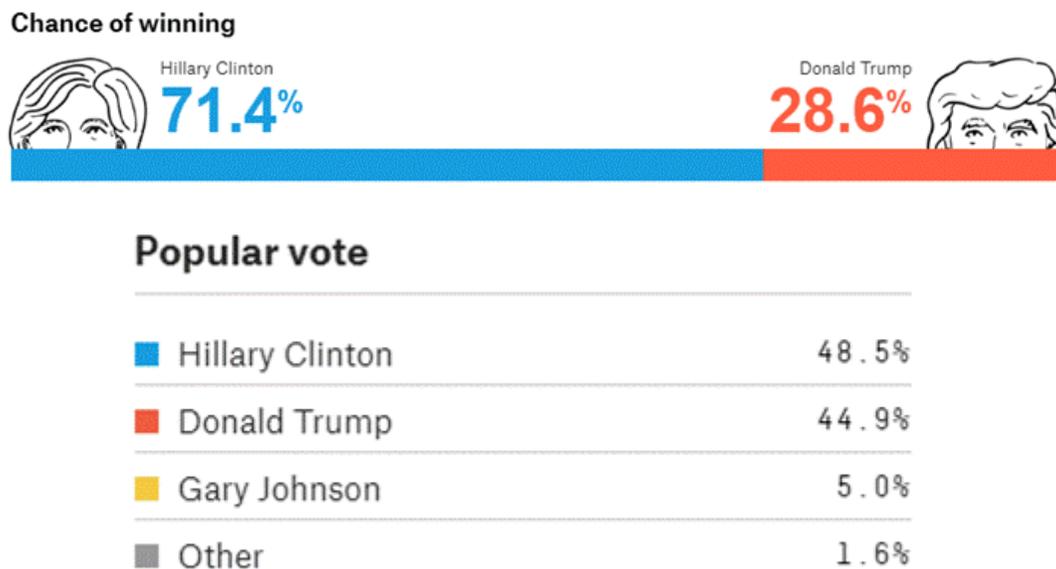
In Bayesian terms, the findings suggest a systematic discrepancy between beliefs about the expected value and tail-mass of the posterior distribution, particularly when considering the more extreme ends of the tails. In Study 6 (and Study A12), we tested directly whether the failure to engage in accurate Bayesian reasoning about chances and margins, as documented in this research, may reflect an active reliance on a mis-specified subjective distribution (e.g., with overestimated tails; Moore et al., 2015; Urminsky, 2014), from which participants' chance and margin beliefs would, in fact, be consistent. We found little support for this interpretation; instead, participants who were more accurate at estimating the prior distribution were no better at the task, and we cannot rationalize participants' estimates based on their subjective priors. In fact, participants' perceptions of their own estimation strategies (S5 and S6) shed little light on their accuracy, beyond the obvious benefit of incorporating the provided forecast into their estimate in some way.

These findings suggest that the supposedly irrelevant choice of format for a forecast can actually have a meaningful impact on people's interpretations of the forecast (as demonstrated by the chance-margin discrepancy) and due to a general forecast-format bias that can even shift attitudes and behaviors. Given that both chance and margin forecast formats consistently yielded biased estimates, but in opposite directions, our evidence suggests that neither format is better or should be favored in communicating information. Instead, the most prudent practice may be to communicate forecasts in both formats simultaneously. Doing so may reduce the potential for forecast-format bias, assuming that people attend to both sources of information and incorporate

both in their subjective beliefs, rather than primarily focusing on and evaluating only one of the cues (Shen & Urminsky, 2013).

The implications for readers of forecasts are similar to the implications associated with other cognitive biases. Readers should be aware that chance forecasts may seem extreme, while margin forecasts may seem to represent small differences. With this awareness, readers can actively adjust their potentially-biased intuition in the reverse direction. Readers may recognize the potential for bias more readily when both forecasts are presented together, triggering conflicting intuitions that the reader must reconcile. While forecasting sites, including *fivethirtyeight.com* (Figure 6), tend to do this already, news reports often present only one format or the other.

Figure 6: Communicating forecasts in both formats (adopted from: *fivethirtyeight.com*)



When only one forecast format is widely available, our findings suggest that a systematic bias may result. Could this bias affect election results? As we demonstrated that the forecast format can affect attitudes, it could plausibly affect intention to vote, as well. In particular, if chance (vs. margin) forecasts leave readers with a stronger sense that the election has already been decided, showing chance forecasts might demotivate voters (Westwood, Messing & Lelkes, 2018).

However, a presidential election is a high-profile event involving substantial news coverage, personal conversations, and other sources of information and preferences beyond forecasts. As a result, many people are likely to have formed behavioral intentions about whether or not they will vote (as well as about other election-related actions) prior to viewing forecasts, so they should have limited sensitivity to manipulated cues (Nickerson & Rogers, 2010), such as forecast format. Furthermore, people choose to vote for a variety of reasons beyond the very low probability that their vote will change the election outcome (see Harder & Krosnick, 2008, for a review).

In considering the potential impact on elections, it is important to take into account that across our studies, the forecast-format bias was weakest when the margin was narrow (e.g., forecasted election results in Study 2). This suggests that forecast format is unlikely to have a causal effect on election outcomes, as the closest elections (which could be affected by a shift in voter turnout) are the least susceptible to the forecast-format bias.

More generally, we tested the potential impact of forecast format on intended election behaviors directly in Study 1, in additional election studies in which people were presented with information about changes in chance or margin forecasts over time (A1, Appendix 1), in waves 4 and 5 in Study 2, and in a replication study (A2, Appendix 2). Forecast format yielded only a

non-significant difference in the self-reported likelihood of voting (Study 1: $M_{\text{chance-forecast-displayed}} = 4.20$ vs. $M_{\text{margin-forecast-displayed}} = 4.46$, $t(223) = 1.47$, $d = -0.20$, $p = .143$) and in intentions for voting and other behaviors in an internal meta-analysis of all relevant studies (Appendix 1). Furthermore, we did not observe a stronger effect of format on behavioral intentions (including voting) among participants living in states where the state-level presidential election was closer, and respondents' votes therefore were more likely to be pivotal (see Appendix 1).

These results do not rule out the possibility that our sample size was not large enough to detect a small but real effect of forecast format on voting intention. However, we expect that downstream effects on behavior are more likely in other contexts, in which people have not already formed behavioral intentions. This is consistent with the effect of forecast format on novel behavior (an unexpected opportunity to bet on an NBA game) that we found in Study 4 (and see additional tests in Appendix 4).

We might also expect to see forecast format play a meaningful role in policy and media discussions of gerrymandering (creating districts to optimize one party's electoral chances). These discussions tend not to involve quantified probabilities of a party winning a gerrymandered district. Instead, both media discussions (Ingraham, 2015) and proposed tests (McGhee, 2014) of gerrymandering tend to focus on the margin-like (and more easily quantified) issue of how the district boundaries impact relative vote shares. When newspaper readers learn that a gerrymandered district has a 60%-40% split between voters of two political parties, for example, they may see the district as more competitive than it is, overestimating the minority party's chance of winning. As a result, readers may erroneously conclude that gerrymandering has less of an impact on election outcomes than it actually does, eroding support for policies

intended to curb gerrymandering. As this example illustrates, people's beliefs about the future may depend on the format of forecasts as much as on the forecasts themselves.

References

- Armstrong JS, Green KC, Graefe A (2015) Golden rule of forecasting: Be conservative. *Journal of Business Research*, 68(8):1717-1731.
- Bowers JS, Davis CJ (2012) Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414.
- Coën A, Desfleurs A, L’Her JF (2009) International evidence on the relative importance of the determinants of earnings forecast accuracy. *Journal of Economics and Business* 61(6):453-471.
- Griffiths TL, Tenenbaum JB (2006) Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773.
- Griffiths TL, Chater N, Norris D, Pouget A (2012) How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, 138(3),:415–422.
- Harder J, Krosnick JA (2008) Why do people vote? A psychological analysis of the causes of voter turnout. *Journal of Social Issues*, 64(3):525–549.
- Hook J (2016, July 17) Hillary Clinton maintains 5-point lead over Donald Trump. *The Wall Street Journal*.
- Ingraham C (2015, March 1) This is the best explanation of gerrymandering you will ever see. *The Washington Post*.
- Jones M, Love BC (2011) Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169–231.

- Kahneman D, Tversky A (1972) Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454.
- Kahneman D, Lovallo D. (1993) Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1):17-31.
- Katz J (2016, July 19) Introducing the Upshot’s presidential prediction model. *The New York Times*.
- Knill DC, Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.
- Kühberger A (1998) The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 75(1):23–55.
- Larrick RP, Soll JB (2008) The MPG illusion. *Science*, 320(5883):1593–1594.
- Levin IP, Gaeth GJ (1988) How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, 15(3):374–378.
- Lieder F, Griffiths T, Goodman N (2012) Burn-in, bias, and the rationality of anchoring. In F. Pereira, CJC Burges, L Bottou, KQ Weinberger (Eds.), *Advances in Neural Information Processing Systems* 25 (pp. 2690–2798). Red Hook, NY: Curran Associates.
- Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.
- McGhee E (2014) Measuring partisan bias in single-member district electoral systems. *Legislative Studies Quarterly*, 39(1):55–85.
- Mochon D, Frederick S (2013) Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes* 122: 69–79

- Moore DA, Carter AB, Yang HH (2015) Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes*, 131:110-120.
- Nickerson DW, Rogers T (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, 21(2):194–199.
- Peachey JA, Schultz DM, Morss R, Roebber PJ, Wood R (2013). How forecasts expressing uncertainty are perceived by UK students. *Weather*, 68(7):176-181.
- Peterson CR, DuCharme WM, Edwards W (1968) Sampling distributions and probability revisions. *Journal of Experimental Psychology*, 76(2):236–243.
- Pouget A, Beck JM, Ma WJ, Latham PE (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178.
- Shen L, Urminsky O (2013) Making sense of nonsense: The visual salience of units determines sensitivity to magnitude. *Psychological Science*, 24(3):297–304.
- Silver N (2019) When we say 70 percent, it really means 70 percent. Retrieved from <https://fivethirtyeight.com/features/when-we-say-70-percent-it-really-means-70-percent/>.
- Spann M, Skiera B (2009) Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55–72.
- Torry H (2019) Economists See U.S. Recession Risk Rising. *Wall Street Journal*, Jan. 10.
- Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2016). Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Science*, 63(2), 497-518.
- Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychological Bulletin*, 76(2):105–110.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Urminsky O (2014) Misestimating probability distributions of repeated events. *Cognitive Science Society Proceedings*.

Westwood, SJ, S Messing, and Y Lelkes (2019) Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *Journal of Politics*, forthcoming.

Wheeler G, Beach LB (1968) Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance*, 3(1):36–46.

Wright FA, Wright AA (2018) How surprising was Trump's victory? Evaluations of the 2016 U.S. presidential election and a new poll aggregation model. *Electoral Studies*, 54:81–89.

**High Chances and Close Margins:
How Equivalent Forecasts Yield Different Beliefs**

ONLINE APPENDIX

1. **Impact of forecast format on election attitudes and intentions (Study A1)**
2. **Forecast format and across-state election forecasts (Studies S1, S2, and A2)**
3. **Effect of forecast format on the interpretation of NFL forecasts (Study A3)**
4. **Effect of forecast format on betting—Summary of all results (Studies 4 and A5–A9)**
5. **Bayesian linear regression for generating forecasts in Study 5**
6. **Additional results for Study 5**
7. **Study A10: Replication of movie genre forecasts study**
8. **Derivation of chance and margin forecasts in Study 6**
9. **Additional results for Study 6**
10. **Study A11: Replication of statistical scenario study**
11. **Study A12: Replication of statistical scenario study**
12. **Additional logistical details for all studies**
13. **Open science materials**
14. **Regression tables**

We briefly report the results of relevant additional measures in the paper's studies (S1–S6) and summarize twelve additional studies (A1–A12), representing all relevant data not reported in the paper. Full data, code, and full stimuli are posted on OSF at <https://osf.io/f4ys6/>.

1. Impact of forecast format on election attitudes and intentions

Study A1 (N = 198 valid completes) was conducted on November 2nd, 2016, six days before the presidential election. AMT participants read about changes over the last 10 days in the election forecast, framed as either a shift in the chance (e.g., from an 85% chance to a 69% chance of Clinton winning) or a shift in the margin (e.g., from a 53% vote share to a 52% vote share for Clinton). Participants rated this change on the same scale as in Study 1 (1 = “*very good news*” to 7 = “*very bad news*”) and were asked several questions about election-related behavioral intentions.

As in Study 1, participants who were shown the change in chance forecasts gave more extreme assessments of the news than those shown the change in margin forecasts ($M_s = 1.90$ vs. 1.36 , $d = 0.46$, $t(196) = 3.03$, $p = .001$). However, the forecast format did not yield a significant difference in self-reported likelihood of voting ($M_{\text{chance-forecast-displayed}} = 4.44$ vs. $M_{\text{margin-forecast-displayed}} = 4.39$, $t(196) = 0.70$, $d = 0.044$, $p = .758$). Tests of the differences between forecast format conditions for all behavioral intention measures in Studies 1 and A1, as well as in Study 2d (Sept. 20, N = 198) and Study 2e (Oct. 24, N = 233) are shown in Table A1, below.

We also coded the closeness of the actual election outcome in each state in order to test whether forecast format impacted behavioral intentions in those states where the election was close and the respondents were perhaps more likely to assume that their vote was pivotal. We found no moderating effect of the vote gap on election attitude ($\beta = -.013$, $p = .231$), optimism ($\beta = -.004$, $p = .599$), worry ($\beta = .002$, $p = .682$), likelihood of voting ($\beta = -.009$, $p = .291$), likelihood of donating or volunteering ($\beta = .010$, $p = .291$), or likelihood of reminding a friend to vote ($\beta = -.018$, $p = .231$). Thus, we rule out the possibility that forecast format has a stronger impact on behavioral intentions in states where the election is close, supporting the view that the forecast-format bias is not likely to have a meaningful impact on a high profile (e.g., U.S. Presidential) election.

Table A1: Effect of forecast format on average election attitudes and behavioral intentions

Measure	Study	Saw Chance	Saw Margin	<i>d</i>	<i>P</i>
Election attitude*	S1	1.76	1.38	.321	.017
	A2	1.90	1.36	.461	.001
	<i>Meta**</i>	<i>1.83</i>	<i>1.37</i>	<i>.388</i>	<i>< .001</i>
Optimistic*	S2d	1.16	1.15	.022	.878
	S2e	1.23	1.10	.180	.174
	<i>Meta**</i>	<i>1.20</i>	<i>1.12</i>	<i>.104</i>	<i>.282</i>
Worried*	S2d	1.15	1.11	.064	.651
	S2e	1.30	1.19	.142	.280
	S1	1.12	1.11	.010	.941
	A2	1.14	1.09	.063	.660
	<i>Meta**</i>	<i>1.18</i>	<i>1.13</i>	<i>.071</i>	<i>.840</i>
Likely to vote	S2d	3.92	4.21	-.207	.146
	S2e	4.16	4.20	-.033	.799
	S1	4.20	4.46	-.196	.143
	A2	4.44	4.39	.044	.758
	<i>Meta**</i>	<i>4.18</i>	<i>4.31</i>	<i>-.100</i>	<i>.136</i>
Donate/volunteer	S2d	1.45	1.63	-.173	.223
	S2e	1.66	1.41	.248	.059
	<i>Meta**</i>	<i>1.57</i>	<i>1.51</i>	<i>.055</i>	<i>.570</i>
Remind friend to vote	S1	3.43	3.52	-.055	.679
	A2	3.77	3.16	.385	.007
	<i>Meta**</i>	<i>3.59</i>	<i>3.35</i>	<i>.154</i>	<i>.115</i>
Post online about candidate	S2	2.18	2.49	-.204	.128
Talk to others about candidate	S2	2.57	2.78	-.139	.299

*coded as difference from midpoint of scale

**In meta-analysis results, p-values are from regression, controlling for study-level fixed effects.

2. Forecast format and across-state election forecasts

Study A2 (N = 155 valid completes) parallels Study 2, but participants viewed forecasts for four different non-identified U.S. states (New Hampshire, Texas, South Carolina, and Arizona) each with a different level of actual support for the candidates. Participants' margin forecast estimates for the leading candidate in the chance-forecast-displayed condition were predicted by the actual forecasts across states (average $\beta = 0.28$, $p < .001$) but were overestimated ($\Delta = +7.02$, $d = 0.89$, $p < .001$). Participants' chance forecast estimates for the leading candidate in the margin-forecast-displayed condition were predicted by the actual forecasts (average $\beta = 3.93$, $p < .001$) but were underestimated ($\Delta = -13.83$, $d = 2.25$, $p < .001$). These findings replicated across all four states for both chance estimates ($ps < .001$) and margin estimates ($ps < .034$).

3. Effect of forecast format on the interpretation of NFL forecasts

In Study A3 (run concurrently with S2c), participants ($N = 223$ valid completes) were shown a chance forecast (71% chance of the Eagles winning over the Browns) or a margin forecast (point-spread: Eagles to win by 6.5 points) for a football game. Participants in the chance-forecast-displayed condition somewhat overestimated the point-spread (a non-significant difference in the predicted direction: $\Delta = +0.22$, $p = .32$), but those in the margin-forecast-displayed condition significantly underestimated the chance of winning ($\Delta = -8.81$, $p < .001$).

In Study A4, participants from the home states of the 2017 Super Bowl teams ($N = 393$ completes) were shown either a chance forecast (61% Patriots vs. Falcons) or a margin forecast (Patriots to win by 3.5 points). Participants in the chance-forecast-displayed condition significantly overestimated the point-spread ($\Delta = +3.61$, $p < .001$), but those in the margin-forecast-displayed condition somewhat underestimated the forecasted chance of winning (a non-significant difference in the predicted direction: $\Delta = -0.94$, $p = .397$). These results generally parallel our findings in Study 3.

4. Effect of forecast format on betting—Summary of all results

In addition to Study 4, we conducted six other studies (total $N = 1,999$) in which we tested whether sports forecast formats impact betting decisions. Table A2 shows the effect of forecast format on (a) how much out of \$5, at 3:1 odds (or out of \$4 with 5:1 odds in Study A4), participants were willing to bet on the team forecasted to win; and (b) the difference in amounts bet on the winning vs. the losing team. Participants bet on a single game in all studies except for Study A9, in which participants allocated the \$5 across bets on two games. All studies except for Study A4 involved NBA basketball games.

Across all the studies in which participants bet on a single game (i.e., omitting A9), they bet more (as a proportion of funds available to bet) on the winning team when shown a chance forecast than when shown a margin (i.e., point-spread) forecast ($M_{\text{chance-forecast-displayed}} = .49$ vs. $M_{\text{margin forecast-display}} = .41$, $t(2118) = 4.50$, $d = 0.19$, $p < .001$). Furthermore, the effect of forecast format was stronger for games in which the forecasted chances were more extreme (interaction $\beta = 0.003$, $t(2116) = 2.32$, $p = .020$). We also find this moderation result as a marginally significant pre-registered interaction test in Study A8. We find no moderation by age, gender, knowledge of the sport, or intentions to watch the game in question (all $ps > .250$).

Table A2: Impact of forecast format on betting decisions

Study	Game (forecasted winner first)	N	p _{win}	M _{chance}	M _{margin}	D	p	Pre-reg
S4	Raptors vs. Hawks, 3/6/18							
NBA	Bet on winner	371	.93	\$3.03	\$2.62	.225	.033	8906
A4	Patriots vs. Falcons, 2/5/17							
NFL	Bet on winner	393	.61	\$2.16	\$2.17	-.005	.963	
	Difference in bets			+\$2.06	+\$2.03	.016	.877	
A5	Raptors vs. Hawks, 3/6/18							
NBA	Bet on winner	225	.93	\$2.96	\$2.71	.124	.353	8892
A6	Trail Blazers vs. Kings, 2/27/18							
NBA	Bet on winner	502	.90	\$1.93	\$1.44	.265	.003	
	Difference in bets			+\$0.39	+\$0.10	.094	.294	8739
A7	Raptors vs. Pistons, 2/26/18							
NBA	Bet on winner	183	.91	\$2.23	\$1.28	.537	< .001	
	Difference in bets			+\$0.98	\$0	.347	.020	8683
A8	Thunder vs. Bulls, 11/15/17							
NBA	Bet on winner	217	.91	\$2.85	\$2.27	.301	.028	
	Difference in bets			+\$2.36	+\$1.70	.285	.037	
	Heat vs. Wizards, 11/15/17							
	Bet on winner	229	.56	\$1.50	\$1.36	.096	.470	
	Difference in bets			+\$0.06	+\$0.17	-.049	.715	
	<i>Difference in difference</i>			+\$2.30	+\$1.53	.172	.072	6762
A9	Thunder vs. Bulls, 11/15/17	204						
NBA	Bet on winner		.91	\$3.40	\$2.98	.226	.108	
	Difference in bets			+\$2.89	+\$2.38	.186	.186	
	Heat vs. Wizards, 11/15/17							
	Bet on winner		.56	\$0.61	\$0.65	-.043	.757	
	Difference in bets			+\$0.13	-\$0.12	.171	.224	
	<i>Difference in difference</i>			+\$2.76	+\$2.50	.086	.540	6778

The pre-reg column includes the AsPredicted pre-registration number and specifies the row with the statistical test that was pre-registered.

5. Bayesian linear regression for generating forecasts in Study 5

We used Bayesian Linear Regression to generate the estimated distribution of outcomes from a sample, and then we computed chance and margin forecasts from that estimated distribution. This approach and the R code for the Bayesian Linear Regression can be found at <https://www.r-bloggers.com/bayesian-linear-regression-analysis-without-tears-r/>.

In Figure A1, we present the code that generated the forecasts. The full code, including the Bayesian Linear Regression estimation routine, can be found in the OSF folder for Study 5.

For each of the 10 pairs of genres, we pulled 200 samples of 15 movies from each genre, for a total of 30 movies in the sample dataset. We then ran a Bayesian Linear Regression, using genre to predict rating, with an improper uniform prior on the regression coefficient. This generated a posterior distribution of the regression coefficient, which estimated the average difference in movie ratings across the two genres. The margin forecast for that sample was simply the median of the distribution, and the chance forecast for that sample was the probability that the coefficient was greater than zero.

This approach has several benefits. First, our forecasts were arguably optimal (under the Bayesian assumptions) given the information in a single sample. Second, because we drew multiple samples, we developed a range of chance-margin forecast pairs that corresponded to the same underlying reality (the true difference between Metacritic movie ratings in the two genres). This allowed us to compare participants' sensitivity to the forecast shown with their sensitivity to the ground truth. Third, because the samples underlying the forecasts were drawn at random, participants viewed a representative sample of forecasts, allowing us to estimate the overall average effect, not just the effect for researcher-chosen test cases.

Figure A1: R code for Bayesian linear regression

```

library(MASS)
# Import the period-level dataset
data_in<-read.csv("Movie_scores.csv")

#Simulation - all pairs
loop<-200
nsamp<-15
result<-matrix(0, nrow = 200*25, ncol = 4)
rx<-1
for (j in 1:4){
  for (k in (j+1):5){
    d1<- data_in[data_in$genrenum==j,]
    d2<- data_in[data_in$genrenum==k,]
    for (i in 1:loop){
      sampdat1<- d1[sample(nrow(d1), nsamp), ]
      sampdat2<- d2[sample(nrow(d2), nsamp), ]
      sampdat <- rbind(sampdat1, sampdat2)
      sampdat$x<-as.numeric(sampdat$genrenum == j)
      lmfit<-lm(score~x,data=sampdat)
      bf<-bayesfit(lmfit,10000)
      outb<-t(apply(bf,2,Bayes.sum))
      result[rx,1]<-j
      result[rx,2]<-k
      result[rx,3]<-outb[2,4]
      result[rx,4]<-1-outb[2,14]
      rx<-rx+1
    }
  }
}
result.df<-as.data.frame(result)
names(result.df)[1]<-"Movie1"
names(result.df)[2]<-"Movie2"
names(result.df)[3]<-"Margin"
names(result.df)[4]<-"Probability"

```

6. Additional results for Study 5

In Study 5, as a conservative test, we included only the 220 participants (out of 400) who gave correct answers in the practice task. As detailed in this section, the results were similar when we included data for all 400 participants. As shown in Table A3, averaging over the sample-specific forecasts for each pair of genres, participants in the margin-forecast-displayed conditions underestimated the probability that the genre favored in the sample also received a higher average rating in the full Metacritic data ($ps < .001$ for all ten genre pairs), with an average underestimation of -25.89 percentage points ($M_{\text{estimated}} = 57.19$ vs. $M_{\text{actual}} = 83.08$, $d = 1.27$, $t(2089) = 47.68$, $p < .001$). Conversely, participants in the chance-forecast-displayed conditions overestimated the margin by which the genre favored in the sample was rated higher in the full Metacritic data ($ps < .001$ for all ten genre pairs), with an average overestimation of 30.64 rating points ($M_{\text{estimated}} = 39.44$ vs. $M_{\text{actual}} = 8.80$, $d = 1.26$, $t(1909) = 45.74$, $p < .001$). A majority of participants (89%) in the chance-forecast-displayed condition overestimated the margin (i.e., difference in ratings), and a majority of participants (89%) in the margin-forecast-displayed condition underestimated the chance (i.e., that the genre favored in the sample received higher ratings).

Table A3: Estimation error for pairs of genres in Study 5 (N = 400)

Genres	Chance Forecast Displayed			Margin Forecast Displayed		
	Est. Margin	Avg. Margin	Error	Est. Chance	Avg. Chance	Error
Action vs. horror	34.3	4.9	+29.5***	52.5	71.6	-19.1***
Adventure vs. crime	34.7	5.4	+29.3***	55.3	75.5	-20.3***
Action vs. adventure	36.8	6.4	+30.4***	55.3	77.0	-21.7***
Action vs. crime	36.5	7.0	+29.5***	55.8	82.2	-26.4***
Adventure vs. horror	39.6	7.3	+32.3***	56.8	81.1	-24.3***
Crime vs. horror	36.2	7.2	+29.0***	56.5	79.1	-22.6***
Biography vs. crime	40.8	9.4	+31.4***	57.5	85.2	-27.7***
Adventure vs. biography	42.3	9.9	+32.4***	56.8	87.6	-30.8***
Action vs. biography	45.5	14.2	+31.3***	62.1	94.6	-32.4***
Biography vs. horror	47.7	16.3	+31.4***	63.4	97.0	-33.6***

A regression predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value that corresponded to the displayed chance forecast ($\beta_{\text{ACTUAL}} = 1.33$, $\eta^2 = 0.075$, $t(1908) = 10.05$, $p < .001$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value that corresponded to the displayed margin forecast ($\beta_{\text{ACTUAL}} = 0.39$, $\eta^2 = 0.068$, $t(2088) = 10.29$, $p < .001$). We again found a significant interaction between estimation condition and actual margin forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -1.52$, partial $\eta^2 = 0.030$, $t(3996) = 9.40$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that while participants' estimates were sensitive to the differences in forecasts across samples, participants nevertheless displayed a chance-margin discrepancy that was more pronounced with more extreme forecasts.

Regardless of whether participants were provided with the genre identities, they underestimated the chance forecast to a similar degree in the margin-forecast-displayed conditions ($\text{Error}_{\text{Genre}} = -27.0$ vs. $\text{Error}_{\text{No-genre}} = -24.6$, $d = .098$, $t(2088) = 0.90$, $p = .373$) and overestimated the margin forecast to a similar degree in the chance-forecast-displayed conditions ($\text{Error}_{\text{Genre}} = 30.8$ vs. $\text{Error}_{\text{No-genre}} = 30.5$, $d = .007$, $t(1908) = 0.06$, $p = .953$).

In the genre-displayed conditions, participants' chance forecast estimates were significantly predicted by the actual, sample-based, forecasted margins ($\beta_{\text{ACTUAL}} = .931$, $t(1107) = 7.63$, $p < .001$) and directionally predicted by the true difference in ratings for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = 1.1$, $t(1107) = 1.69$, $p = .09$). Likewise, participants' margin forecast estimates were significantly predicted by the actual, sample-based, forecasted margin difference ($\beta_{\text{ACTUAL}} = 1.03$, $t(967) = 5.29$, $p < .001$) but not by the true differences in ratings for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = 0.59$, $t(967) = 0.71$, $p = .481$).

Lastly, we analyzed participants' open-ended descriptions of the strategies they used. Overall, 54.8% of participants indicated that they had used the forecast provided, and 26% of participants indicated that their strategy involved scaling higher forecasts into higher estimates, with no significant difference between the margin-display and chance-display conditions.

The specific strategies, as shown in Table A4, differed across conditions ($\chi^2 = 21.3$, $df = 10$, $p = .019$). The most common answer (43.3%) described a non-specific, intuition-based estimate, often using the provided forecast; many participants, though, did not describe a strategy

at all (26.3%). Relatively few described completely ineffective strategies, such as random guessing (3.5%) or adopting the provided forecast as the estimate (1.8%).

Table A4: Participants' self-reported decision strategies in Study 5 (N = 400)

	All	Chance Forecast Displayed	Margin Forecast Displayed
Nonspecific intuition	43.3%	47.6%	39.2%
Anchoring and adjusting	7.8%	11.4%	12.1%
Using standard deviation or statistical distribution	7.0%	9.9%	4.3%
Using mathematics and the forecast	6.0%	3.1%	8.6%
Taking a random guess	3.5%	3.7%	3.3%
Assuming chances and share were the same	1.8%	2.1%	1.4%
Ranking the possible outcomes	0.3%	0%	0.5%
Estimating near the average	0.5%	0%	1.0%
Using own opinion of genres	3.0%	2.6%	3.3%
Other	0.8%	1.6%	0%
No strategy provided	26.3%	24.1%	28.2%

Overall, the results suggest that some strategies were less effective than others. Participants who mentioned using the forecast ($r = -.26, p < .001$) or scaling up their estimates with larger forecasts ($r = -.27, p < .001$) had lower average absolute errors. Furthermore, participants who mentioned using the standard deviation or a statistical distribution had lower average absolute errors ($r = -.21, p < .001$), although this was also the case for those who described a non-specific intuition ($r = -.15, p = .002$). That said, the chance-margin discrepancy persisted regardless of strategy used. In the margin-forecast-displayed conditions, the chance forecast was underestimated by 22.8 percentage points among participants who mentioned using the forecast ($p < .001$) and by 18.9 percentage points among those who mentioned using the standard deviation or distribution ($p < .001$). Similarly, in the chance-forecast-displayed conditions, the margin forecast was overestimated by 24.9 percentage points among participants who mentioned using the forecast ($p < .001$) and by 10.5 percentage points among those who mentioned using the standard deviation or distribution ($p < .001$).

7. Study A10: Replication of movie genre forecasts study

This study used the same Metacritic rating movie genre forecasts as in Study 5, and participants were randomly assigned to either the chance-forecast-displayed or margin-forecast-displayed condition. However, participants did not do the training tasks from in Study 5, and each participant was asked to make an estimate for only one genre pair. Five participants were selected at random to be paid a linear incentive for the accuracy of their estimates, with a maximum accuracy bonus of \$5 per person.

Out of a total of 502 valid participants, 66 gave estimates in the opposite direction of the provided forecast (e.g., estimating that Genre B had higher ratings in the full Metacritic data even though the provided forecast indicated that Genre A had higher ratings in the sample). As a conservative test, we report results excluding these participants to ensure that our results are not driven by a subset of participants who misunderstood the forecasts. The results using all participants revealed even higher levels of systematic error and are discussed at the end of this section.

As shown in Table A5, averaging over the sample-specific forecasts for each pair of genres, participants in the margin-forecast-displayed condition underestimated the chance that the genre favored in the sample was higher rated in the population ($ps < .01$ for four of ten genre pairs), with an average underestimation of -9.36 percentage points ($M_{\text{estimated}} = 72.58$ vs. $M_{\text{actual}} = 81.94$, $d = 0.62$, $t(179) = 6.84$, $p < .001$). Conversely, participants in the chance-forecast-displayed condition overestimated the margin by which the genre favored in the sample was rated higher in the population ($ps < .001$ for all ten genre pairs), with an average overestimation of 29.91 rating points ($M_{\text{estimated}} = 38.58$ vs. $M_{\text{actual}} = 8.68$, $d = 1.28$, $t(255) = 15.8$, $p < .001$). A majority of participants (88%) in the chance-forecast-displayed condition overestimated the expected difference in ratings, and a majority of participants (71%) in the margin-forecast-displayed condition underestimated the chance that the genre favored in the sample was rated higher in the population. As in the prior studies, both forecast formats yielded biased estimates, but the estimates were biased in opposite directions.

Table A5: Estimation error for pairs of genres in Study A10 (N = 436)

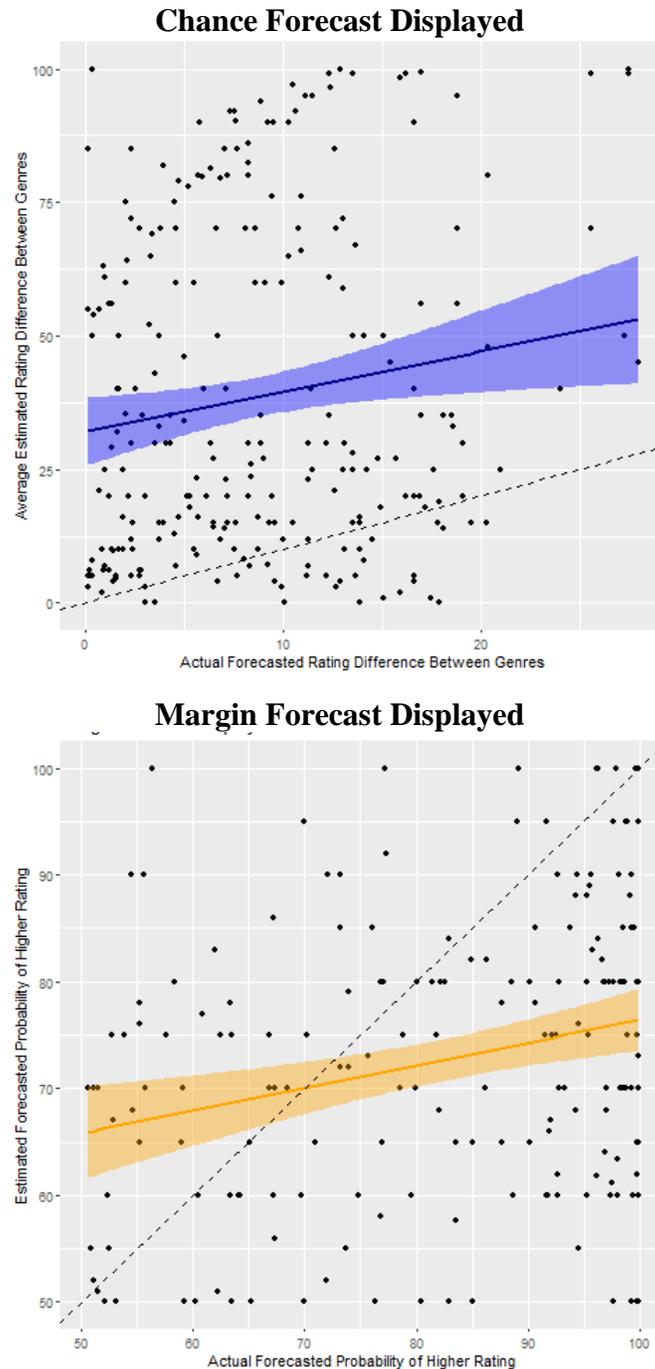
Genres	Chance Forecast Displayed			Margin Forecast Displayed		
	Est. Margin	Avg. Margin	Error	Est. Chance	Avg. Chance	Error
Action vs. adventure	25.2	5.7	+19.5***	71.1	76.4	-5.2
Action vs. biography	49.2	14.4	+34.8***	73.1	96.3	-23.2***
Action vs. crime	38.9	7.6	+31.3***	74.7	76.4	-1.7
Action vs. horror	35.7	3.3	+32.4***	71.6	74.6	-3.0
Adventure vs. biography	47.2	10.3	+36.9***	73.7	86.9	-13.2**
Adventure vs. crime	37.3	5.2	+32.1***	70.8	75.1	-4.3
Adventure vs. horror	36.3	6.2	+30.1***	74.6	76.7	-2.1
Biography vs. crime	39.8	8.4	+31.5***	66.0	83.0	-17.0***
Biography vs. horror	40.4	15.6	+24.8***	76.0	97.1	-21.1***
Crime vs. horror	34.2	8.7	+25.6***	73.8	82.1	-8.3

As shown in Figure A2, participants overestimated the margin relative to the actual forecast and generally underestimated the chance (although forecasted chances near 50% were overestimated).

A regression predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value that corresponded to the provided chance forecast ($\beta_{\text{ACTUAL}} = 0.75$, $\eta^2 = 0.024$, $t(254) = 2.51$, $p = .013$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value that corresponded to the provided margin forecast ($\beta_{\text{ACTUAL}} = 0.21$, $\eta^2 = 0.064$, $t(178) = 3.48$, $p < .001$). We again found a significant interaction between estimation condition and actual margin forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.74$, partial $\eta^2 = 0.055$, $t(432) = 5.01$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that while participants' estimates were sensitive to the differences in forecasts across samples, participants nevertheless displayed a forecast-format bias that was more pronounced with more extreme forecasts.

Figure A2: Estimation results based on different forecast formats in Study A10

The dashed black lines represent the actual forecasts derived from the sample, while the colored lines indicate the participants' estimates (chance in orange, margin in blue), made after viewing the other forecast format; the confidence intervals are shaded. The estimate of each individual participant appears as a black diamond.



Participants' chance forecast estimates were significantly predicted by the actual, sample-based, forecasted margin difference ($\beta_{\text{ACTUAL}} = 0.42$, $t(177) = 2.25$, $p = .025$) but not by the true differences for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = 1.37$, $t(177) = 1.57$, $p = .118$). Likewise, participants' margin forecast estimates were significantly predicted by the actual, sample-based, forecasted margin difference ($\beta_{\text{ACTUAL}} = 0.91$, $t(253) = 2.85$, $p = .005$) but not by the true differences for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = -2.30$, $t(253) = 1.40$, $p = .161$). Thus, the observed misestimation of forecasts cannot be explained by the possibility that participants adjusted inaccurate forecasts (e.g., those based on non-representative samples) to reflect the true differences in movie ratings more accurately.

The results are similar when including data for all 502 participants. As shown in Table A6, averaging over the sample-specific forecasts for each pair of genres, participants in the margin-forecast-displayed condition underestimated the probability that the genre favored in the sample was higher rated in the population ($ps < .01$ for four of ten genre pairs), with an average underestimation of -22.71 percentage points ($M_{\text{estimated}} = 60.08$ vs. $M_{\text{actual}} = 82.79$, $d = 1.05$, $t(237) = 11.69$, $p < .001$). Conversely, participants in the chance-forecast-displayed condition overestimated the margin by which the genre favored in the sample was rated higher in the population ($ps < .001$ for all ten genre pairs), with an average overestimation of 28.05 rating points ($M_{\text{estimated}} = 36.76$ vs. $M_{\text{actual}} = 8.71$, $d = 1.16$, $t(263) = 14.30$, $p < .001$). A majority of participants (85%) in the chance-forecast-displayed condition overestimated the expected difference in ratings, and a majority of participants (78%) in the margin-forecast-displayed condition underestimated the chance that the genre favored in the sample was rated higher in the population .

Table A6: Estimation error for pairs of genres in Study A10 (N = 502)

Genres	Chance Forecast Displayed			Margin Forecast Displayed		
	Est. Margin	Avg. Margin	Error	Est. Chance	Avg. Chance	Error
Action vs. adventure	25.2	5.7	+19.5***	67.9	76.2	-8.3
Action vs. biography	36.6	13.7	+22.9**	52.3	97.1	-44.8***
Action vs. crime	38.9	7.6	+31.3***	55.7	77.7	-22.1*
Action vs. horror	35.7	3.3	+32.4***	56.7	74.1	-17.5*
Adventure vs. biography	45.1	10.3	+34.8***	65.7	87.6	-21.9***
Adventure vs. crime	37.3	5.2	+32.1***	59.8	74.8	-15.0**
Adventure vs. horror	36.3	6.2	+30.1***	66.7	77.2	-10.6
Biography vs. crime	38.0	8.4	+29.6***	54.5	85.2	-30.7***
Biography vs. horror	40.4	15.6	+24.8***	59.8	96.9	-37.1***
Crime vs. horror	32.2	8.6	+23.6***	59.8	82.0	-22.2***

A regression predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value that corresponded to the provided chance forecast ($\beta_{\text{ACTUAL}} = 0.66$, $\eta^2 = 0.017$, $t(262) = 2.14$, $p = .033$). However, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition did not find significant sensitivity to the actual chance forecast value that corresponded to the provided margin forecast ($\beta_{\text{ACTUAL}} = 0.07$, $\eta^2 = 0.002$, $t(236) = 0.68$, $p < .495$). We again found a significant interaction between estimation condition and true value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.86$, partial $\eta^2 = 0.053$, $t(498) = 5.29$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that the forecast-format bias was more pronounced with more extreme forecasts.

Participants' chance forecast estimates were not significantly predicted by either the actual, sample-based, forecasted margin difference ($\beta_{\text{ACTUAL}} = -0.23$, $t(235) = 0.75$, $p = .451$) or the true difference for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = 1.81$, $t(235) = 1.21$, $p = .226$). However, participants' margin forecast estimates were significantly predicted by the actual, sample-based, forecasted margin difference ($\beta_{\text{ACTUAL}} = 0.84$, $t(261) = 2.56$, $p = .011$) but

not by the true difference for the genre pair in the full Metacritic data ($\beta_{\text{TRUE}} = -2.67$, $t(261) = 1.57$, $p = .117$). Thus, the observed misestimation of forecasts is not explained by adjusting inaccurate forecasts (e.g., due to non-representative samples) to be more accurate of the true genre differences in movie ratings.

These results are similar to those in Study 5 and confirm that people make incompatible chance/margin predictions even when using a representative set of uniform-prior Bayesian forecasts.

8. Derivation of chance and margin forecasts in Study 6

For Studies 6, A11, and A12 we designed the scenarios such that a Bayesian would be able to compute the correct forecast from the information provided, as follows. Let $n \sim U[0, N]$ be the unknown number of red marbles in a jar ($N-n$ marbles are green, for a total of N). A sample of S marbles are drawn with replacement, yielding some number s of red marbles in the sample. The expected number of red marbles in the jar is then the expected value of the posterior distribution, conditional on the sample outcome s . The chance that the majority of marbles in the jar are red can be calculated from the mass of the posterior distribution at or above $N/2$.

Defining i as a given outcome (the number of red marbles in the sample), we want to compute $E(n | s = i)$ and $P(n > N/2 | s = i)$. We use Bayes' Rule:

$$P(n = j | s = i) = \frac{P(s = i | n = j)P(n = j)}{P(s = i)}$$

Since n is drawn from a uniform distribution, for all integer values of j in $[0, N]$:

$$P(n = j) = \frac{1}{N + 1}$$

By the definition of the binomial distribution:

$$P(s = i | n = j) = \binom{S}{i} \left(\frac{j}{N}\right)^i \left(1 - \frac{j}{N}\right)^{S-i}$$

Lastly, we can sum over these terms to compute the denominator:

$$P(s = i) = \sum_{k=0}^N P(s = i | n = k)P(n = k)$$

The expected number of red marbles for the margin forecast is therefore given by:

$$E(n | s = i) = \sum_{r=0}^N r * P(n = r | s = i)$$

The probability that at least half of the marbles are red, for the chance forecast, is:

$$P(n \geq \frac{N}{2} | s = i) = \sum_{r=N/2}^N P(n = r | s = i)$$

We computed the relevant quantities for $N = 99$ and $S = 20$ in R. The stimuli shown to participants were based on sample outcomes of $i = 11$ through $i = 14$ out of 20 marbles being red.

To compute the relevant quantities for $N = 99$ and $S = 20$, we used the following R script:

```
# Binomial Simulation
parameters
Ncap<-99
Scap<-20

# initialize vectors
expnumred<-rep(0, Scap+1)
probmaxred<-rep(0, Scap+1)
ProbS<-rep(0, Scap+1)

# compute baseline probabilities of outcomes
for(s in 0:Scap){
  for(k in 0:Ncap)
  {
    ProbS[s+1] <- ProbS[s+1]+dbinom(s, size=Scap, prob=k/Ncap )*(1/(Ncap+1))
  }
}

# compute expected outcome and probability of majority red
for(i in 0:Scap){
  for(r in 0:Ncap)
  {
    # E(n|s=i)
    expnumred[i+1] <- expnumred[i+1]+r*dbinom(i, size=Scap,
prob=r/Ncap)*(1/(Ncap+1))/ProbS[i+1]
    # P(n>=N/2|s=i)
    if (r>=Ncap/2) {probmaxred[i+1]=probmaxred[i+1]+dbinom(i, size=Scap,
prob=r/Ncap)*(1/(Ncap+1))/ProbS[i+1]
    }
  }
}

# display results
expnumred
probmaxred
ProbS
```

This yields the following estimates that we used as stimuli for the study:

Table A7: Statistical inferences based on samples

Number of red marbles i in sample, out of 20	11	12	13	14
Expected number of red marbles, $E(n s = i)$	54.0	58.5	63.0	67.5
Expected margin	9	18	27	36
Probability that most marbles are red, $P(n > N/2 s = i)$.6682	.8084	.9055	.9609

9. Additional details for Study 6

The following literal descriptions of decision strategies were used in Study 6 in the chance-display conditions.

Table A8: Strategy descriptions in chance-forecast-displayed conditions (Study 6)

Strategy	Survey description
Nonspecific intuition	I used my intuition and can't explain exactly how my estimate came to me (3)
Bayesian reasoning	I thought about jars with different numbers of red and green marbles, and then estimated how many more red marbles to expect by averaging based on the likelihood of samples of 20 from each jar giving a [81%/96%] chance of more red (8)
Using mathematics and the forecast	I used my knowledge of mathematics to compute how many more red marbles to expect based on the [81%/96%] chance of more red marbles. (2)
Random guessing	I just guessed randomly (1)
Assuming chances and share were the same	I assumed that [81%/96%] chance meant [81%/96%] of the marbles were red and then calculated the difference in the number of red and green marbles (5)
Anchoring and adjusting (correct direction)	I started with the [81%/96%] chance and adjusted downwards to a smaller number (4)
Mentally simulating the sampling process	I thought about jars with different numbers of red and green marbles and then imagined taking marbles from the jar 20 times and then picked the imaginary jar that felt like it would give me more red marbles [81%/96%] of the time. (6)
Ranking the possible outcomes	I thought about all the possible jars with different numbers of red and green marbles and picked a number of red marbles so that [81%/96%] of jars would have fewer than that amount (7)
Anchoring and adjusting (wrong direction)	I started with the [81%/96%] chance and adjusted upwards to a larger number (9)
Other	Other (please describe) (10)

The following literal descriptions of decision strategies were used in Study 6 in the margin-display conditions.

Table A9: Strategy descriptions in margin-display conditions (Study 6)

Strategy	Survey description
Nonspecific intuition	I used my intuition and can't explain exactly how my estimate came to me (3)
Bayesian reasoning	I thought about jars with different numbers of red and green marbles, and then estimated the probability of more red marbles to expect by combining the probability based on the likelihood of samples of 20 from each jar giving a forecast of [18/36] more red (8)
Using mathematics and the forecast	I used my knowledge of mathematics to compute the probability of more red marbles based on the forecast of [18/36] more red marbles. (2)
Random guessing	I just guessed randomly (1)
Assuming chances and share were the same	I started with [18/36] more red marbles out of 99 and then calculated what percent of the marbles were red (5)
Anchoring and adjusting (correct direction)	I started with the [18/36] more red marbles and adjusted upwards to a larger number (4)
Mentally simulating the sampling process	I thought about jars with different numbers of red and green marbles and then imagined taking marbles from the jar 20 times and then picked the probability based on the imaginary jar that felt like it would give me about [20%/35%] more red marbles (6)
Ranking the possible outcomes	I thought about all the possible jars with different numbers of red and green marbles and then estimated what percent of jars would have fewer than [19/37] more red than green marbles (7)
Anchoring and adjusting (wrong direction)	I started with the [18/36] more red marbles and adjusted downwards to a smaller number (9)
Other	Other (please describe) (10)

Additional results.

Two levels of forecasts (corresponding to a sample with either 12 or 14 red marbles) were tested in Study 6, between-subjects. A regression predicting participants' margin forecast estimates in the chance-forecast-displayed condition did not find significant sensitivity to the actual margin forecast value ($\beta_{\text{ACTUAL}} = -0.09$, $\eta^2 = 0.002$, $t(291) = 0.82$, $p = .416$). A regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found only marginally significant sensitivity to the true chance forecast value ($\beta_{\text{ACTUAL}} = 0.37$, $\eta^2 = 0.013$, $t(283) = 1.96$, $p = .052$). We did not find a significant interaction between estimation condition and actual chance forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.33$, partial $\eta^2 = 0.003$, $t(574) = 1.33$, $p = .185$) in a follow-up regression predicting estimation error.

We believe that the lack of evidence that the forecast-format bias was more pronounced with more extreme actual forecasts was in part due to the relatively low power and low evaluability in this between-subjects design. Supporting this interpretation, we consistently find both sensitivity to true forecasts and a greater chance-margin discrepancy for more extreme actual forecasts in our supplemental studies that used the same paradigm but tested multiple forecasts in within-subjects designs (see Studies A11 and A12, below).

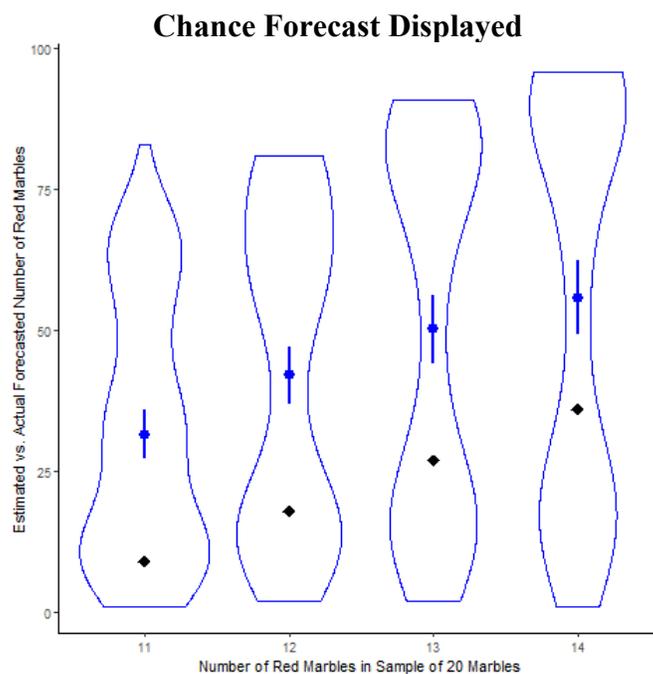
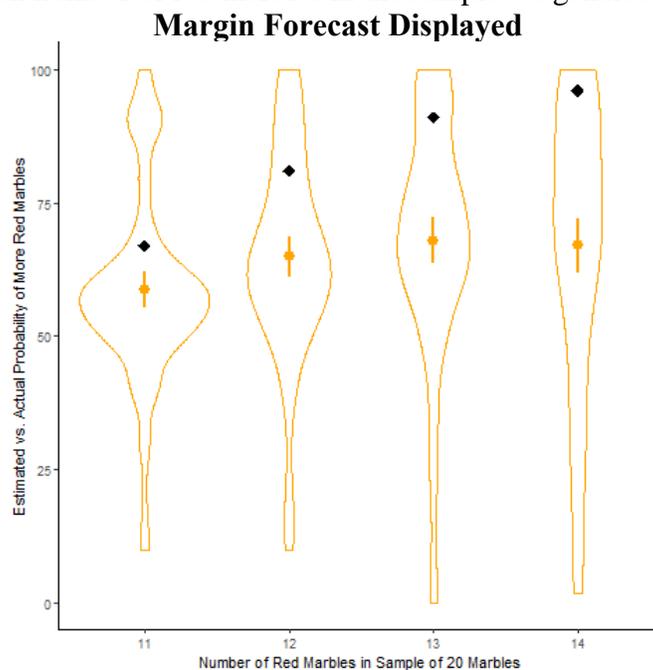
10. Study A11: Replication of Statistical Scenario (Study 6)

Study A11 (N = 218 valid AMT participants) was a replication of Study 6, with each participant making four judgments instead of just one. As in Study 6, participants were randomly assigned to either the chance-forecast-displayed or margin-forecast-displayed condition. In Study A11, participants were shown four scenarios corresponding to four sample results (as described in Appendix Section 8), all showing one type of forecast, and then estimated the corresponding other format of forecast. We counterbalanced the order in which we presented the forecast information for red and green marbles. Each participant was paid a linear incentive for the accuracy of their estimates, with a maximum accuracy bonus of \$1 per person.

As shown in Figure A3, participants in the margin-forecast-displayed condition significantly underestimated the chance, in all four scenarios, that more marbles were red than green (all $ps < .001$), with an average underestimation of -19.15 percentage points ($M_{\text{estimated}} = 64.60$ vs. $M_{\text{actual}} = 83.75$, $d = 1.26$, $t(105) = 12.99$, $p < .001$). Conversely, participants in the chance-forecast-displayed conditions significantly overestimated the margin by which red marbles outnumbered green marbles in all four scenarios (all $ps < .001$), with an average overestimation of 22.44 marbles ($M_{\text{estimated}} = 44.94$ vs. $M_{\text{actual}} = 22.50$, $d = 0.78$, $t(111) = 8.25$, $p < .001$). As in the prior studies, both forecast formats yielded biased estimates, but the estimates were biased in opposite directions.

Figure A3: Estimation results based on different forecast formats in Study A11

Colored dots indicate participants' mean estimates (chance estimates in orange; margin estimates in blue), made after viewing the other forecast format; error bars represent the confidence intervals, and the colored plots show the distribution of estimates. Black diamonds represent the actual forecasts for each number of red marbles in the sample we generated from the population.



A regression with clustered standard errors predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the actual margin forecast value ($\beta_{\text{ACTUAL}} = 0.90$, $\eta^2 = 0.083$, $t(446) = 12.44$, $p < .001$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the actual chance forecast value ($\beta_{\text{ACTUAL}} = 0.31$, $\eta^2 = 0.024$, $t(422) = 3.97$, $p < .001$). As in the prior studies, the chance-margin discrepancy was pervasive across participants—63% of participants in the chance-forecast-displayed condition overestimated the margin of red vs. green marbles, averaged across the four tasks, while 88% of participants in the margin-forecast-displayed condition underestimated the chance that red marbles outnumbered green marbles. This discrepancy was not explained by anchoring (i.e., adopting the provided forecast as the estimate). We again found a significant interaction between estimation condition and the actual chance forecast value ($\beta_{\text{FORMAT} \times \text{ACTUAL}} = -0.62$, partial $\eta^2 = 0.017$, $t(868) = 6.07$, $p < .001$) in a follow-up regression predicting estimation error, indicating a robust forecast-format bias that was more pronounced with more extreme actual forecasts.

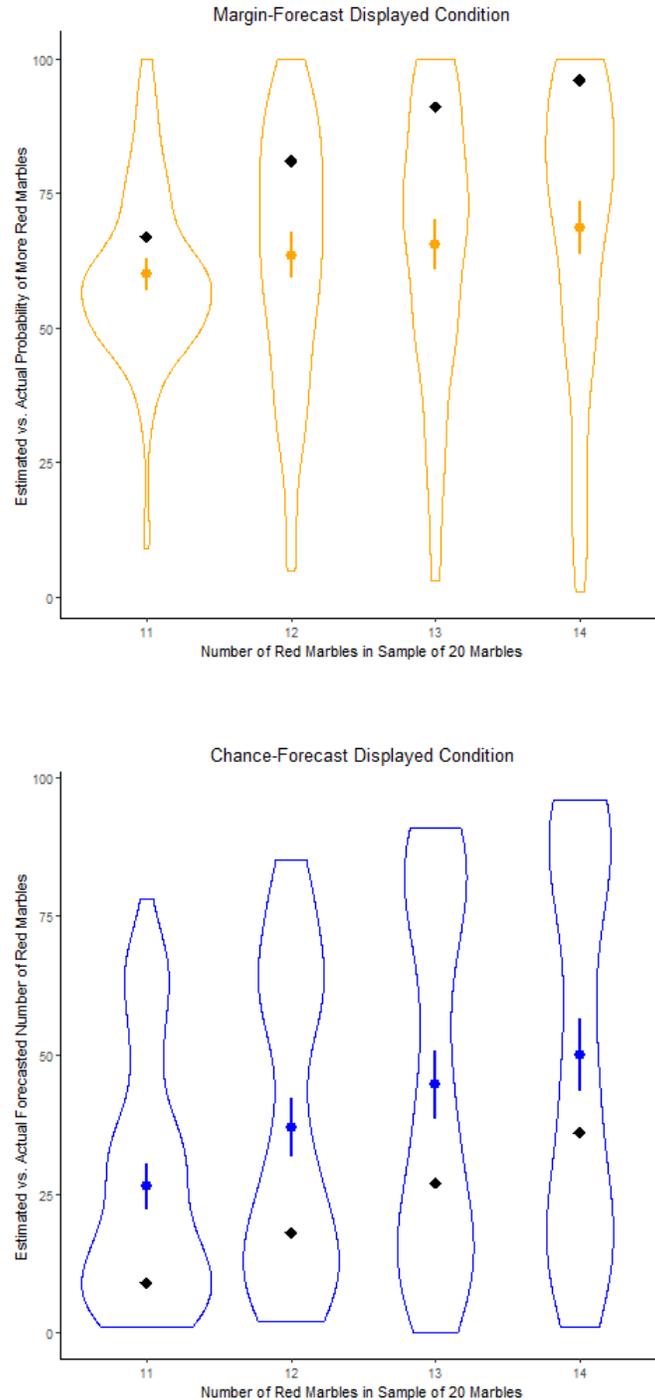
11. Study A12: Replication of Statistical Scenario (Study 6)

Study A12 (N = 211 valid AMT participants) used most of the same methods as Study A11 but did not counterbalance the display order for red and green marbles. Participants in Study A12 also completed a probability distribution estimation task (as in Study 6). In the distribution task, participants were asked to imagine that 10 marbles were sampled with replacement from a jar containing 50 red and 50 green marbles. Using an adjustable histogram, participants estimated the joint probability of the 11 possible outcomes (from 0 red and 10 green to 10 red and 0 green marbles).

As shown in Figure A4, participants in the margin-forecast-displayed condition significantly underestimated the chance, in all four scenarios, that more marbles were red than green (all $ps < .001$), with an average underestimation of -19.37 percentage points ($M_{\text{estimated}} = 64.38$ vs. $M_{\text{actual}} = 83.75$, $d = 1.65$, $t(109) = 12.26$, $p < .001$). Conversely, participants in the chance-forecast-displayed conditions significantly overestimated the margin by which red marbles outnumbered green marbles in all four scenarios (all $ps < .001$), with an average overestimation of 16.97 marbles ($M_{\text{estimated}} = 39.43$ vs. $M_{\text{actual}} = 22.47$, $d = 0.84$, $t(100) = 6.29$, $p < .001$).

Figure A4: Estimation results based on different forecast formats in Study A12

Colored dots indicate participants' mean estimates (chance estimates in orange; margin estimates in blue), made after viewing the other forecast format; error bars represent the confidence intervals, and the colored plots show the distribution of estimates. Black diamonds represent the actual forecasts for each number of red marbles in the sample we generated from the population.



The chance-margin discrepancy was pervasive across participants—62% of participants in the chance-forecast-displayed condition overestimated the margin of red vs. green marbles, averaged across the four tasks, while 84% of participants in the margin-forecast-displayed condition underestimated the chance that red marbles outnumbered green marbles—and this discrepancy was not explained by anchoring (i.e., adopting the provided forecast as the estimate). As in the prior studies, both forecast formats yielded biased estimates, but the estimates were biased in opposite directions.

A regression with clustered standard errors predicting participants' margin forecast estimates in the chance-forecast-displayed condition found significant sensitivity to the true margin forecast value ($\beta_{\text{TRUE}} = 0.88$, $\eta^2 = 0.087$, $t(401) = 11.89$, $p < .001$). Likewise, a regression predicting participants' chance forecast estimates in the margin-forecast-displayed condition found significant sensitivity to the true chance forecast value ($\beta_{\text{TRUE}} = 0.28$, $\eta^2 = 0.018$, $t(438) = 3.60$, $p < .001$). We again found a significant interaction between estimation condition and true value ($\beta_{\text{FORMAT} \times \text{TRUE}} = 0.64$, partial $\eta^2 = 0.019$, $t(839) = 6.21$, $p < .001$) in a follow-up regression predicting estimation error. These results suggest that while participants' estimates were sensitive to the differences in forecasts across samples, participants nevertheless displayed a forecast-format bias that was more pronounced with more extreme forecasts.

Finally, we tested whether participants with more accurate statistical intuition (as determined by the binomial sampling distribution scenario) are less prone to the forecast-format bias. We found no significant interaction between estimation condition and absolute error in estimating binomial distributions ($\beta_{\text{FORMAT} \times \text{BIN_ERR}} = -0.07$, partial $\eta^2 < 0.001$, $t(839) = 0.32$, $p = .748$) in a regression predicting estimation error. We found a marginally significant interaction between estimation condition and, specifically, the overestimation of the tails of the binomial distribution ($\beta_{\text{FORMAT} \times \text{TAILS}} = 0.36$, partial $\eta^2 = 0.011$, $t(839) = 1.89$, $p = .060$) in a regression predicting estimation error. However, in this regression, the predicted effect of estimation condition for participants with *no* error in estimating the tails was still highly significant ($\beta_{\text{FORMAT}} = 41.18$, $t(839) = 9.83$, $p < .001$). Thus, these analyses are consistent with the conclusion in Study 6: that the observed bias is not explained by participants' generally mistaken beliefs about the shape of the binomial distribution.

12. Additional logistical details for all studies

- Participants were assigned to either S2b, A1, or another study that measured estimates in the absence of any forecasts.
- Participants were assigned to either S2c, A3, or a marble-scenario. Participants were assigned to either S2d, A3, or a marble-scenario. The marble scenarios yielded similar results but were not included because they contrasted sample observations with chance forecasts (as opposed to margin forecasts with chance forecasts, as in S5).
- Participants were assigned to either S2f or another study that tested the effects of participants anchoring on numbers equivalent to the forecasts or the actual values, but not identified as such.

13. Open science materials

The pre-registrations associated with the studies can be found online as follows:

Study 3 (AsPredicted #6391): <http://aspredicted.org/blind.php?x=ba38f8>

Study 4 (AsPredicted #8906): <http://aspredicted.org/blind.php?x=5qw89v>

Study 5 (As Predicted #33246): <http://aspredicted.org/blind.php?x=36zh25>

Study 6 (AsPredicted #24681): <http://aspredicted.org/blind.php?x=mk9en6>

Study A5 (AsPredicted #8892): <http://aspredicted.org/blind.php?x=y9pd7c>

Study A6 (AsPredicted #8739): <http://aspredicted.org/blind.php?x=jx9kt7>

Study A7 (AsPredicted #8683): <http://aspredicted.org/blind.php?x=7sv8xw>

Study A8 (AsPredicted #6762): <http://aspredicted.org/blind.php?x=w8p4z9>

Study A9 (AsPredicted #6778): <http://aspredicted.org/blind.php?x=qp7xn7>

Study A11 (AsPredicted #6686): <http://aspredicted.org/blind.php?x=gq8hm3>

Study A12 (AsPredicted #23210): <https://aspredicted.org/blind.php?x=9z94gi>

In addition, data files, survey materials, pre-registrations, and analysis code (in R) for the studies are available on OSF (<https://osf.io/f4ys6/>).

14. Regression tables

Variable definitions:

ActualMargin = actual margin forecast

ActualChance = actual chance forecast

SawMargin = 1 in margin-display-condition, 0 otherwise

TrueDifference = true difference in ratings between genres for full Metacritic universe

BinomialAbsError = absolute error in estimation of binomial distribution

BinomialTails = error in tails of estimated binomial distribution (positive = overestimated)

Table A10: Regression predicting estimate (Study 2, chance-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	-54.21	26.39	-2.05	.040
ActualMargin	2.18	0.50	4.35	< .001

Table A11: Regression predicting estimate (Study 2, margin-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	40.56	4.22	9.62	< .001
ActualChance	0.26	0.06	4.54	< .001

Table A12: Regression predicting estimation error (Study 2)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	-1.76	4.12	-0.43	.670
ActualChance	0.13	0.05	2.39	.017
SawMargin	42.31	5.80	7.30	< .001
ActualChance x SawMargin	-0.88	0.08	-11.34	< .001

Table A13: Regression predicting estimate (Study 3, chance-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	5.60	0.53	10.63	< .001
ActualMargin	1.66	0.14	11.51	< .001

Table A14: Regression predicting estimate (Study 3, margin-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	29.01	3.96	7.33	< .001
ActualChance	0.44	0.06	7.15	< .001

Table A15: Regression predicting estimation error (Study 3)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	-4.52	2.16	-2.09	.037
ActualChance	0.20	0.04	4.64	< .001
SawMargin	33.53	4.51	7.44	< .001
ActualChance x SawMargin	-0.76	0.07	-10.12	< .001

Table A16: Regression predicting estimate (Study 5, chance-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	14.45	1.92	7.53	< .001
ActualMargin	1.53	0.16	9.39	< .001

Table A17: Regression predicting estimate (Study 5, margin-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	23.90	2.42	9.87	< .001
ActualChance	0.46	0.03	13.97	< .001

Table A18: Regression predicting estimation error (Study 5)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	14.45	1.91	7.56	< .001
ActualMargin	0.53	0.16	3.28	.001
SawMargin	-26.91	2.15	-12.50	< .001
ActualMargin x SawMargin	-1.56	0.19	-8.39	< .001

Table A19: Regression predicting estimate (Study 5, chance-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	16.45	2.97	5.54	< .001
ActualMargin	1.43	0.24	5.93	< .001
TrueDifference	0.63	0.81	0.78	.436

Table A20: Regression predicting estimate (Study 5, margin-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	51.83	1.42	36.49	< .001
ActualChance	1.21	0.11	11.47	< .001
TrueDifference	0.63	0.52	1.22	.224

Table A21: Regression predicting estimate (Study 6, chance-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	27.75	5.29	5.24	< .001
ActualMargin	0.37	0.19	1.96	.052

Table A22: Regression predicting estimate (Study 6, margin-display condition)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	80.62	9.50	8.49	< .001
ActualChance	-0.09	0.11	-0.82	.416

Table A23: Regression predicting estimation error, moderation by actual chance (Study 6)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	77.82	15.62	4.98	< .001
ActualChance	-0.76	0.18	-4.30	< .001
SawMargin	2.80	21.95	0.13	.899
ActualChance x SawMargin	-0.33	0.25	-1.33	.185

Table A24: Regression predicting estimation error, moderation by binomial absolute error (Study 6)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	10.47	2.63	3.97	< .001
BinomialAbsError	0.01	0.06	0.19	.847
SawMargin	-24.11	3.75	-6.43	< .001
BinomialAbsError x SawMargin	-0.07	0.09	-0.74	.463

Table A25: Regression predicting estimation error, moderation by estimated tails of binomial distribution (Study 6)

	Estimate	SE	<i>t</i>	<i>p</i>
(Intercept)	8.35	2.38	3.51	< .001
BinomialTails	0.06	0.05	1.32	.188
SawMargin	-23.52	3.33	-7.07	< .001
BinomialTails x SawMargin	-0.07	0.07	-1.06	.292

