EXPOSURE TO SIMILAR VS. DIVERSE PERSPECTIVES IN FORECASTING TOURNAMENTS ON HUMAN WELFARE AND SOCIETAL CHANGE

By

Kevin Chen

An Undergraduate Thesis submitted as part of the

WHARTON RESEARCH SCHOLARS

Faculty Advisor:

Philip E. Tetlock, Annenberg University Professor, Wharton School of Business

Cory J. Clark, Behavioral Scientist, University of Pennsylvania

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MAY 2023

Abstract

What are the intra-individual benefits to participating in forecasting tournaments? This paper presents the findings of an empirical study testing how exposure to similar vs. diverse perspectives in forecasting tournaments affects distributions of predictions and estimated ranges, judgmental accuracy, belief updating, and affective polarization. Everyday participants were randomly assigned to one of two conditions—exposure to similar vs. diverse perspectives—and made individual forecasts and rationales for four domains, reviewed similar vs. diverse team forecasts and rationales, optionally updated their individual forecasts and rationales in light of team information, and completed other individual difference assessments. Across domains, (i) there were no significant main effect of perspective, main effect of stage, nor interaction on judgmental accuracy; (ii) participants exposed to similar perspectives updated their predictions more often than did those exposed to diverse perspectives; (iii) participants exposed to similar perspectives rated conservatives more warmly than did those exposed to diverse perspectives (with nonsignificant but directionally consistent results for ratings toward liberals and moderates).

Keywords: forecasting tournaments, diverse perspectives, belief updating, affective polarization

Exposure to Similar vs. Diverse Perspectives in

Forecasting Tournaments on Human Welfare and Societal Change

What are the intra-individual benefits to participating in forecasting tournaments? Forecasting tournaments are competitions in which participants attempt to make the most accurate predictions about specific future outcomes. A key interest of research involves identifying the most skilled and accurate forecasters ("superforecasters") and using their insights toward making more calibrated and refined predictions about future outcomes (Mellers et al., 2015; Tetlock & Gardner, 2015). The structure of forecasting tournaments may vary, but typically they involve several rounds of forecasts with the possibility to work individually and in teams, obtain accuracy feedback, and update one's predictions in light of new information (Tetlock et al., 2014).

The broader literature on forecasting ranges from predicting the outcomes of sports (Wunderlich & Memmert, 2020), forecasting the winner of political elections (Williams & Reade, 2015), estimating the performance of companies and industries (Kapoor & Wilde, 2022), predicting the progress of artificial intelligence (Gruetzemacher et al., 2021), and judging the trajectory of pandemics (Davies & Ferris, 2022). Forecasting is relevant to business, politics, and society, and its applications are wide-ranging in making informed decisions about the future.

Within this broader umbrella of forecasting literature, researchers are also interested in the potential social and psychological benefits to participating in forecasting tournaments. Forecasting tournaments can not only be an engaging way for participants to learn more about different topics, but they can also challenge participants' assumptions and beliefs as they reckon with the (in)accuracy of their predictions. Importantly, this may help individuals to develop their critical thinking skills and become more open-minded and flexible in their judgments and beliefs.

Tetlock et al. (2014) discuss these potential benefits, including how forecasting tournaments can contribute to depolarizing scientific and policy debates, and Mellers et al. (2019) demonstrate how participating in forecasting tournaments can decrease polarization and enhance charitable attributions to political adversaries.

This paper seeks to contribute to this growing literature on the intra-individual benefits to participating in forecasting tournaments by investigating how exposure to similar vs. diverse perspectives in forecasting tournaments affects the distributions of predictions and estimated ranges, judgmental accuracy, belief updating, and affective polarization.

Distributions of Predictions and Estimated Ranges and Judgmental Accuracy

There is an extensive literature on diversity and its positive and negative influences on individual and team performance (for a review, see van Knippenberg & Mell, 2016). In terms of forecasting—a complex and open-ended task—research suggests that diversity may have positive effects on performance requiring the integration and synthesization of heterogeneous perspectives and information (Guillaume et al., 2015; van Knippenberg & Mell, 2016). Diversity of thought can be beneficial in forecasting tournaments because different individuals may have different ways of analyzing data and coming up with different interpretations of information, leading to a more comprehensive view of a situation. Having a diverse group of individuals sharing their thoughts and perspectives may help to counteract biases that may be present in the data or in the forecasting process. Overall, informational diversity may be an important factor contributing to the performance of forecasters in forecasting tournaments.

Research by Pescetelli et al. (2021) supports the idea that group diversity (based on a large multi-trait space) can improve individual and aggregate forecaster accuracy judged on Brier scoring. Furthermore, they show that larger groups (~25 people) benefitted from more group

diversity while higher levels of group diversity had a negative impact on the accuracy of smaller teams (~5 people). Exploratory data analyzing the distributions of forecasts revealed no effects of diversity (or group size) on the variability in forecasts and decreased dispersion between initial (solo) forecasts and final (team) forecasts, reflecting greater consensus after group interaction (Pescetelli et al., 2021).

Belief Updating

In the context of forecasting tournaments, belief updating refers to the process of updating a forecast based on new information. Research on belief updating in forecasting tournaments has shown that the most accurate forecasters tend to make frequent, small updates, whereas less accurate forecasters are prone to confirm their initial judgments or make infrequent, large revisions. Furthermore, high-frequency updaters scored higher on open-mindedness, which suggests a relationship between belief updating behavior and intellectual humility (Atanasov et al., 2020).

Exposure to diverse perspectives may promote belief updating because it can provide individuals with a wider range of information and perspectives to consider when evaluating their beliefs. This may encourage participants to be more open-minded and flexible in their thinking and to be more willing to revise their beliefs in light of new evidence. On the other hand, exposure to similar perspectives may also be useful in forecasting tournaments as it may provide a common frame of reference that can help participants better understand and evaluate the information that is held in consensus. An individual forecaster who diverges from their team's homogeneous forecasts and rationales may be more likely to update their beliefs to conform with the group, or they may decisively double-down on their prior beliefs. Drawing on Pescetelli et al.'s (2021) finding that group diversity can improve forecasting accuracy, it may follow that

exposure to diverse forecasts and rationales leads to more incremental belief updating behavior (a tendency of more accurate forecasters), whereas exposure to homogeneous beliefs leads to infrequent, but on average larger updates (a tendency of less accurate forecasters). However, it is ultimately unclear how exposure to similar vs. diverse perspectives in forecasting tournaments might affect belief updating.

Related to belief updating, the epistemic virtue of intellectual humility broadly refers to the willingness to admit that one's beliefs may be wrong and the openness to revising or changing them in light of new evidence (Whitcomb et al., 2017; Leary, 2018). Intellectual humility is considered an important virtue in many fields, particularly in academia and public policy, because it encourages individuals to consider alternative viewpoints and to update their beliefs in light of new evidence. Intellectual humility has been shown to be associated with open-minded thinking (Krumrei-Mancuso et al., 2020; Leary et al., 2017; Porter & Schumann, 2018) as well as less affective polarization (Krumrei-Mancuso & Newman, 2020).

Participation in forecasting tournaments may serve to foster intellectual humility in a number of ways. First, the act of making forecasts and being judged on their accuracy can help individuals recognize their own limitations and biases. By having their forecasts tested against reality and potentially being proven wrong, participants may become more aware of the ways in which their preconceptions and assumptions influence their thinking. Additionally, exposure to other participants' forecasts and rationales may help individuals become more open to new perspectives, as seeing how others approach the same problem and come to different conclusions can challenge participants' beliefs and encourage them to think more critically about the issues at hand. Overall, participating in forecasting tournaments and being exposed to a variety of

different perspectives may help to foster intellectual humility as individuals are confronted with their own limitations and become more receptive to new ideas and ways of thinking.

Affective Polarization

Affective polarization is broadly defined as animosity between parties (Iyengar et al., 2019). Some of the most pressing social issues facing the United States include political polarization, which may be exacerbated by shortcomings in epistemic virtues such as open-mindedness and intellectual humility. Prior research has shown negative correlations between measures of intellectual humility and affective polarization, suggesting that interventions boosting intellectual humility might also decrease affective polarization (Bowes et al., 2020; Bowes et al., 2021). Forecasting tournaments have been shown to decrease polarization and enhance charitable attributions to political adversaries (Mellers et al., 2019).

Exposure to diverse perspectives can be beneficial to reducing affective polarization because it can provide individuals with a wider range of information to consider when evaluating their beliefs. This may encourage individuals to recognize that there are many different viewpoints on any given topic and to be more open to hearing and understanding those viewpoints. On the other hand, exposure to similar perspectives can contribute to reducing affective polarization by providing a shared experience to help individuals find common ground with others. While participating in forecasting tournaments can reduce polarization (Mellers et al., 2019), it is difficult to say how exposure to similar vs. diverse perspectives within forecasting tournaments might mediate participants' affective polarization.

Thus, by exploring how exposure to similar vs. diverse perspectives in forecasting tournaments affects the distributions of predictions and estimated ranges, judgmental accuracy,

belief updating, and affective polarization, this paper seeks to contribute to the growing literature studying how participating in forecasting tournaments might confer benefits to participants.

Context

The dataset I analyze is from the first of a three-year Forecasting Tournament on Human Welfare and Societal Change. The longitudinal study investigates how forecasting tournaments can foster judgmental accuracy, intellectual humility, and open-mindedness in debates over human progress, drawing on participation from subject matter experts (SMEs) and individuals with established track records of superior accuracy in forecasting events ("superforecasters"; Mellers et al., 2015; Tetlock & Gardner, 2015), and the attentive public. Over the course of three years (2022 through 2024), the study will host two parallel forecasting tournaments—one with SMEs and superforecasters and another with everyday people—to investigate epistemic virtues and their development over time.

Specifically, this paper focuses on data from year one (2022) of the everyday people tournament, a dataset of 518 everyday participants and their predictions, estimated ranges, and individual differences. Everyday participants were recruited from CloudResearch and Lucid through the completion of an online recruitment survey in which they completed individual difference assessments. Then, respondents with fully completed, generally coherent responses were invited to the everyday people tournament, with additional consideration of sample representativeness (e.g., gender, education, socioeconomic/political views). Over the span of five recruitment waves (multiple waves were needed to reach a sufficient sample size), invited participants were randomly assigned into either exposure to similar or diverse forecasts and rationales.

In the forecasting tournament, participants completed an online survey in which, for each domain, they were provided a graph of the historical data and were prompted to:

(i) provide their individual forecasts—2 short-term and 1 long-term prediction and estimated range—and rationales for each domain (i.e., "individual stage");

(ii) review pseudo team forecasts—predictions and estimated ranges—and rationales, vote for convincing team rationales, and optionally updated individual forecasts and rationales (i.e., "team stage");

(iii) complete other individual difference assessments.

Figure 1 provides a screenshot of the tournament user interface where participants were prompted to enter their predictions and estimated ranges. Participants made three predictions and three estimated ranges for four domains critical to human welfare and societal change, a total of twelve forecasts (prediction and estimated range) per participant in the individual stage with the option to update forecasts in the team stage after being exposed to either similar or diverse perspectives.

The four critical domains participants made forecasts for were nominated by SMEs prior to the everyday people tournament, and were climate (e.g., average global CO2 concentrations), public health (e.g., global infant mortality rate), the economy (e.g., poverty rate in the United States), and global peace/war (e.g., global occurrence of non-state conflicts). In this paper, I restrict my analysis to three of the four domains—climate, economy, and peace/war—due to data discrepancies in the domain of public health complicating the analysis on judgmental accuracy.

After participants completed the forecasting tournament (individual and team stages), they were directed to complete a survey that collected data on individual differences. Individual difference assessments spanned measures for accountability, intellectual humility (Alfano et al.,

2017; Krumrei-Mancuso & Rouse, 2015), general knowledge, fluid intelligence (e.g., Raven's Matrices; Arthur & Day, 1994; Bilker et al., 2012), personality (e.g., HEXACO; Ashton et al., 2014), primals (Clifton et al., 2019), ideology, and affective polarization (Lavrakas, 2008), among others. For the purposes of this paper, my analysis only pertains to the measurements collected for affective polarization. Combined, the tournament and individual difference assessments were intended to last no longer than 90 minutes.

Method

Exposure to Similar vs. Diverse Perspectives

The predictor variable of interest was participants' exposure to similar vs. diverse forecasts and rationales during the team stage of the forecasting tournament. After providing individual forecasts and rationales for each domain, participants were ostensibly assigned to a team of five other everyday participant forecasters. In this team stage, participants were prompted to review their team members' individual forecasts and rationales, vote for two convincing rationales, and update their individual forecasts and/or rationales if they so desired.

In reality, team members' forecasts and rationales were predetermined, adapted from real responses made by SMEs in their parallel tournament. From SME responses, I selected similar and diverse forecasts and rationales, tweaked them for clarity and to accentuate their similarity/diversity, and then for rationales, I manually coded for unique themes mentioned (e.g., COVID, Russia-Ukraine conflict, population growth, carbon capture, etc.). Between conditions within domains, total word count and sentence count were controlled, and the ratios of unique themes (similar:diverse) were less than or equal to .25 for all domains. A pilot-test was run in which an independent sample of participants from Lucid were given the ten proposed rationales per domain and asked to label the five most similar rationales as '1's and the five least similar

rationales as ‘0’s. As shown in Table 1, there was sufficient evidence (all *p*s < .001) to conclude a difference between the similar and diverse rationales which were later used in the everyday people tournament.

Distributions of Predictions and Estimated Ranges

Everyday participants made 1-year, 2-year, and 20-year predictions as well as highest and lowest estimates per time horizon for each domain, once in the individual stage with a chance to update in the team stage after exposure to similar or diverse perspectives. For forecast distributions, I investigated how exposure to similar vs. diverse perspectives affected participants’ predictions and estimated ranges—calculated by subtracting lowest estimates from corresponding highest estimates—for each domain. Extreme outliers for predictions and estimated ranges, defined as observations more than three times the interquartile range beyond the first and third quartiles, were excluded in the analysis for each domain and time horizon.

For predictions, I computed nine mixed-design ANOVAs (three time horizons over three domains), with a fixed effects factor being exposure to perspectives (similar vs. diverse) and a random effects factor being the stage (individual vs. team). Similarly for estimated ranges, I computed nine mixed-design ANOVAs (three time horizons over three domains), with a fixed effects factor being exposure to perspectives (similar vs. diverse) and a random effects factor being the stage (individual vs. team).

Judgmental Accuracy

Accuracy was scored for resolved 1-year predictions by the absolute value of the difference between actual outcomes and participant predictions. For each domain, I computed a mixed-design ANOVA, with a fixed effects factor being exposure to perspectives (similar vs. diverse) and a random effects factor being the stage (individual vs. team).

Belief Updating

In the team stage of the forecasting tournament, participants had the option to update their predictions in light of information from team members' similar or diverse perspectives. Over three domains and three time horizons, this amounted to a total of nine opportunities per participant to update their predictions (without considering opportunities to update high/low estimates). Identical with the analyses on predictions and estimated ranges, observations with extreme outliers for predictions and estimated ranges for each domain and time horizon were excluded in the analysis for belief updating.

Affective Polarization

From the individual difference assessments following the forecasting tournament, feeling thermometers—survey items that ask respondents to rate their feelings toward a person, group, or issue on a numerical scale—were used to measure affective polarization. After participating in the forecasting tournament, participants were prompted to rate their feelings toward liberals, moderates, and conservatives on a scale of 0 to 100, with 0 representing the lowest possible level of positive feelings and 100 representing the highest possible level of positive feelings (Lavrakas, 2008).

To investigate how exposure to similar vs. diverse perspectives affected feelings of warmth toward liberals, moderates, and conservatives, I computed two-sample *t*-tests between participants exposed to similar perspectives and those exposed to diverse perspectives. No outliers were excluded.

Results

Distributions of Predictions and Estimated Ranges

Figures 1 through 3 present distributions of predictions per domain, and Table 2 presents summary statistics. For predictions in the climate domain, there was:

(i) no significant main effect of perspective ($F(1, 472) = 0.20$, $p = .66$), main effect of stage ($F(1, 472) = 1.44$, $p = .23$), nor interaction ($F(1, 472) = 0.09$, $p = .76$) on 1-year climate predictions;

(ii) no significant main effect of perspective ($F(1, 476) = 0.23$, $p = .63$), main effect of stage ($F(1, 476) = 3.07$, $p = .08$), nor interaction ($F(1, 476) = 0.28$, $p = .60$) on 2-year climate predictions;

(iii) no significant main effect of perspective ($F(1, 477) = 0.28$, $p = .60$), main effect of stage ($F(1, 477) = 0.24$, $p = .62$), nor interaction ($F(1, 477) = 2.78$, $p = .10$) on 20-year climate predictions.

For predictions in the economy domain, there was:

(i) no significant main effect of perspective ($F(1, 492) = 0.39$, $p = .54$), but a significant main effect of stage ($F(1, 492) = 12.29$, $p < .05$) and interaction ($F(1, 492) = 11.46$, $p < .05$) on 1-year economy predictions;

(ii) no significant main effect of perspective ($F(1, 496) = 0.01$, $p = .91$) nor main effect of stage ($F(1, 496) = 1.38$, $p < .24$), but a significant interaction ($F(1, 496) = 7.83$, $p < .05$) on 2-year economy predictions;

(iii) no significant main effect of perspective ($F(1, 493) = 0.91$, $p = .34$) nor main effect of stage ($F(1, 493) = 0.88$, $p < .35$), but a significant interaction ($F(1, 493) = 8.25$, $p < .05$) on 20-year economy predictions.

For predictions in the peace/war domain, there was:

(i) no significant main effect of perspective ($F(1, 437) = 3.01$, $p = .08$), main effect of stage ($F(1, 437) = 0.19$, $p = .67$), nor interaction ($F(1, 437) = 1.54$, $p = .22$) on 1-year peace/war predictions;

(ii) no significant main effect of perspective ($F(1, 472) = 0.64$, $p = .42$), main effect of stage ($F(1, 472) = 1.17$, $p = .28$), nor interaction ($F(1, 472) = 0.38$, $p = .54$) on 2-year peace/war predictions;

(iii) no significant main effect of perspective ($F(1, 490) = 1.20$, $p = .27$) nor interaction ($F(1, 490) = 0.16$, $p = .68$), but a significant main effect of stage ($F(1, 490) = 25.79$, $p < .05$) on 20-year peace/war predictions.

Figures 4 through 6 present distributions of estimated ranges per domain, and Table 3 presents summary statistics. For estimated ranges in the climate domain, there was:

(i) no significant main effect of perspective ($F(1, 472) = 1.84$, $p = .18$) nor interaction ($F(1, 472) = 0.01$, $p = .93$), but a significant main effect of stage ($F(1, 472) = 8.60$, $p < .05$) on 1-year climate ranges;

(ii) no significant main effect of perspective ($F(1, 476) = 0.04$, $p = .84$) nor interaction ($F(1, 476) = 0.32$, $p = .57$), but a significant main effect of stage ($F(1, 476) = 8.87$, $p < .05$) on 2-year climate ranges;

(iii) no significant main effect of perspective ($F(1, 477) = 0.75$, $p = .39$), but a significant main effect of stage ($F(1, 477) = 6.96$, $p < .05$) and interaction ($F(1, 477) = 6.18$, $p < .05$) on 20-year climate ranges.

For estimated ranges in the economy domain, there was:

(i) no significant main effect of perspective ($F(1, 492) = 0.27$, $p = .61$), main effect of stage ($F(1, 492) = 1.41$, $p = .24$), nor interaction ($F(1, 492) = 0.65$, $p = .42$) on 1-year economy ranges;

(ii) no significant main effect of perspective ($F(1, 496) = 0.00$, $p = .96$), main effect of stage ($F(1, 496) = 0.98$, $p = .32$), nor interaction ($F(1, 496) = 2.02$, $p = .16$) on 2-year economy ranges;

(iii) no significant main effect of perspective ($F(1, 493) = 0.30$, $p = .58$), main effect of stage ($F(1, 493) = 3.68$, $p = .06$), nor interaction ($F(1, 493) = 0.92$, $p = .34$) on 20-year economy ranges.

For estimated ranges in the peace/war domain, there was:

(i) no significant main effect of perspective ($F(1, 437) = 0.52$ $p = .47$), main effect of stage ($F(1, 437) = 0.70$, $p = .40$), nor interaction ($F(1, 437) = 0.92$, $p = .34$) on 1-year peace/war ranges;

(ii) no significant main effect of perspective ($F(1, 472) = 1.00$, $p = .32$), main effect of stage ($F(1, 472) = 0.51$, $p = .40$), nor interaction ($F(1, 472) = 2.54$, $p = .11$) on 2-year peace/war ranges;

(iii) no significant main effect of perspective ($F(1, 490) = 0.22$, $p = .64$), but a significant main effect of stage ($F(1, 490) = 11.84$, $p < .05$) and interaction ($F(1, 490) = 4.39$, $p < .05$) on 20-year peace/war ranges.

Judgmental Accuracy

Table 4 presents summary statistics of accuracy by domain. In the climate domain, there was no significant main effect of perspective ($F(1, 472) = 0.35$, $p = .56$), main effect of stage

($F(1, 472) = 0.35$, $p = .56$), nor interaction ($F(1, 472) = 1.42$, $p = .23$) on 1-year accuracy. In the economy domain, there was no significant main effect of perspective ($F(1, 492) = 0.09$, $p = .76$), main effect of stage ($F(1, 492) = 0.93$, $p = .34$), nor interaction ($F(1, 492) = 0.26$, $p = .61$) on 1-year accuracy. In the peace/war domain, there was no significant main effect of perspective ($F(1, 437) = 1.52$, $p = .22$), main effect of stage ($F(1, 437) = 0.02$, $p = .88$), nor interaction ($F(1, 437) = 0.20$, $p = .65$) on 1-year accuracy.

Belief Updating

Across domains, there was a significant difference in the percentage of updates between participants exposed to similar vs. diverse perspectives ($X^2(1) = 13.16$, $p < .05$), where participants exposed to similar perspectives updated their predictions more often than did those exposed to diverse perspectives.

Table 5 breaks down percentages of belief updating by domain. In the climate domain, there was not a significant difference in the percentage of updates ($X^2(1) = 1.23$, $p = .27$). In the economy domain, there was a significant difference in the percentage of updates ($X^2(1) = 11.72$, $p < .05$). In the peace/war domain, there was not a significant difference in the percentage of updates ($X^2(1) = 1.89$, $p = .17$). Within all three domains, there was a directional difference where participants exposed to similar perspectives updated their predictions at a higher rate than did those exposed to diverse perspectives.

Correlations Between Judgmental Accuracy and Belief Updating

Examining Pearson correlations between judgmental accuracy in the team stage and belief updating, correlation coefficients were not significantly different from zero in the climate ($t(472) = -0.36$, $p = .72$), economy ($t(492) = -0.64$, $p = 0.52$), nor peace/war domains ($t(437) = 0.86$, $p = .39$).

Affective Polarization

While there was not a significant difference between participants exposed to similar vs. diverse perspectives regarding their mean feeling thermometer scores toward liberals ($t(514.8) = -1.39$, $p = .16$) and toward moderates ($t(515.6) = -1.78$, $p = .08$), there was a directional difference where participants exposed to similar perspectives rated liberals and moderates more warmly than did those exposed to diverse perspectives. There was a significant difference between participants exposed to similar vs. diverse perspectives regarding their mean feeling thermometer scores toward conservatives ($t(513.1) = -2.12$, $p < .05$), where participants exposed to similar perspectives rated conservatives more warmly than did those exposed to diverse perspectives (i.e., directionally consistent with ratings toward liberals and moderates). Figure 7 presents box plots of feeling thermometer scores, and Table 6 presents summary statistics.

Discussion

Distributions of Predictions and Ranges

Among everyday participants, exposure to similar vs. diverse perspectives did not impact the predictions and estimated ranges across short and long-term time horizons in the climate, economy, and peace/war domains. However, we do observe that, as the time-horizon expands from 1- to 2- to 20-year forecasts, the standard deviations of participants' predictions consistently increase within each domain, reflecting that longer time horizons are more uncertain and generally more difficult to forecast. Similarly, the means and standard deviations of participants' estimated ranges also consistently increase with time horizon within each domain, reflecting less participant confidence and greater variability in their confidence the longer the prediction's time horizon.

One thing of particular note regarding the distribution of predictions is that, while predictions in the economy and peace/war domain had a mostly normal distribution, 1- and 2-year predictions in the climate domain had a stark bimodal distribution at prediction values of 420 and 425 for 1-year predictions and 420 and 430 for 2-year predictions across perspectives. I speculate this is in part due to participants anchoring on the y-axis labels found in the presented historical data, as the maximum y-axis label was 425, and labels were separated in increments of 25, which may be suggestive of increments of 5.

Regarding the distribution of estimated ranges, it was notable that estimated ranges across time horizons and domains all exhibited a right skew. Considering that, on average, estimated ranges for 1-year forecasts were 6.84, 2.97, and 15.0 and judgmental accuracy scores were 4.13, 1.01, and 4.27 in the climate, economy, and peace/war domains respectively, assuming estimated ranges were symmetric about participants' predictions, it suggests that participants may have been slightly overconfident in the climate domain, slightly under-confident in the economy domain, and moderately under-confident in the peace/war domain—though this is simply conjecture for now and can be tested in a follow-up study.

Judgmental Accuracy and Belief Updating

Among everyday participants, exposure to similar vs. diverse perspectives did not impact judgmental accuracy for 1-year predictions in the climate, economy, and peace/war domains, and there was a significant difference in belief updating behavior across domains. Within the three domains, there was a directional difference where participants exposed to similar perspectives updated their predictions at a higher rate than did those exposed to diverse perspectives. This suggests that the everyday participants may have been more likely to update their beliefs when the team members' perspectives were similar, perhaps in part to conform with the group,

whereas when exposed to a wider range of information and perspectives, participants tended to not revise their beliefs in light of new evidence. Furthermore, overall belief updating rates were low for all everyday participants, which may suggest overall poor judgmental accuracy (Atanasov et al., 2020), though there was no significant linear relationship between judgmental accuracy in the team stage and belief updating within domains.

Affective Polarization

Considering there was at least a directional difference where participants exposed to similar perspectives rated liberals, moderates, and conservatives more warmly than did those exposed to diverse perspectives (conservatives being a significant difference), this may suggest that exposure to similar perspectives is more beneficial to reducing affective polarization by providing a shared sense of experience and common ground.

One notable caveat is that, toward liberals and moderates, average ratings were very close to or significantly above 50 while average ratings toward conservatives were 46.3 for participants exposed to similar perspectives and 40.4 for those exposed to diverse perspectives, both considerably below 50. Given that measurements were only collected post-tournament, it is difficult to interpret whether feeling thermometer scores went up, stayed the same, or dropped.

In addition, there was high variability in scores toward all political groups, particularly liberals and conservatives. For further study, it would be interesting to investigate how feeling thermometer scores differ from prior to post-tournament participation, or across multiple years of tournament participation, as is the case with the Forecasting Tournament on Human Welfare and Societal Change.

Further Study

Given the scope of the forecasting tournament to three specific domains of human welfare and societal change, questions of generalizability to other prediction domains are valid. However, within these domains, there are a plethora of possible follow-up studies relevant to the existing data and future data.

Firstly, data was collected on a wide variety of individual difference variables, and identifying relationships between other individual differences besides affective polarization with tournament metrics like predictions, accuracy, belief updating, etc. may be valuable, especially those closely associated with open-mindedness and intellectual humility. Secondly, given a parallel tournament was conducted with SMEs and superforecasters, it would be interesting to examine the similarities and differences between SMEs, superforecasters, and everyday people regarding their predictions, estimated ranges, accuracy, belief updating, etc.

Beyond generating knowledge about how judgmental accuracy, belief updating, and affective polarization relate to each other, this research also addresses some of the most pressing social issues facing the United States and other nations. Political polarization is at an all-time high, and attachments to political identities are now stronger than attachments to gender, race, religion, and ethnicity (Iyengar et al., 2018). Forecasting tournaments—that is, experiencing first-hand the challenges of formulating accurate beliefs and the fallibility of one's perceptions—may be the key to recognizing that adversaries, like oneself, are simply doing their best to understand the world in all its complexity.

Further research can help translate empirical findings into practical building blocks for targeted interventions. By learning which mechanisms (e.g., accountability, perspective-taking, performance feedback, perception of a situation as adversarial vs. cooperative, etc.) account for the most variance in promoting intra-individual change in affective polarization and other

variables of importance, practitioners can construct customized interventions oriented toward fostering judgmental accuracy and/or reducing polarization. Follow up studies will be important to further examine the intra-individual benefits associated with participating in forecasting tournaments—benefits that may include, but are certainly not limited to, judgmental accuracy, belief updating, and affective polarization.

References

Alfano, M., Iurino, K., Stey, P., Robinson, B., Christen, M., Yu, F., & Lapsley, D. (2017). Development and validation of a multi-dimensional measure of intellectual humility. In R. E. Tractenberg (Ed.), PLOS ONE (Vol. 12, Issue 8, p. e0182950). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0182950

Arthur, W., JR., & Day, D. V. (1994). Development of a Short form for the Raven Advanced Progressive Matrices Test. In Educational and Psychological Measurement (Vol. 54, Issue 2, pp. 394–403). SAGE Publications. https://doi.org/10.1177/0013164494054002013

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors. In Personality and Social Psychology Review (Vol. 18, Issue 2, pp. 139–152). SAGE Publications. https://doi.org/10.1177/1088868314523838

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. Organizational Behavior and Human Decision Processes (Vol. 160, pp. 19–35). Elsevier BV. https://doi.org/10.1016/j.obhdp.2020.02.001

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test. In Assessment (Vol. 19, Issue 3, pp. 354–369). SAGE Publications. https://doi.org/10.1177/1073191112446655

Bowes, S. M., Blanchard, M. C., Costello, T. H., Abramowitz, A. I., & Lilienfeld, S. O. (2020). Intellectual humility and between-party animus: Implications for affective polarization in

two community samples. Journal of Research in Personality, 88, 103992. https://doi.org/10.1016/j.jrp.2020.103992

Bowes, S. M., Costello, T. H., Lee, C., McElroy-Heltzel, S., Davis, D. E., & Lilienfeld, S. O. (2021). Stepping Outside the Echo Chamber: Is Intellectual Humility Associated With Less Political Myside Bias? Personality and Social Psychology Bulletin, 014616722199761. https://doi.org/10.1177/0146167221997619

Clifton, J. D. W., Baker, J. D., Park, C. L., Yaden, D. B., Clifton, A. B. W., Terni, P., Miller, J. L., Zeng, G., Giorgi, S., Schwartz, H. A., & Seligman, M. E. P. (2019). Primal world beliefs. In Psychological Assessment (Vol. 31, Issue 1, pp. 82–99). American Psychological Association (APA). https://doi.org/10.1037/pas0000639

Davies, N., & Ferris, S. (2022). Human judgement forecasting tournaments: A feasibility study based on the COVID-19 pandemic with public health practitioners in England. In Public Health in Practice (Vol. 3, p. 100260). Elsevier BV. https://doi.org/10.1016/j.puhip.2022.100260

Gruetzemacher, R., Dorner, F. E., Bernaola-Alvarez, N., Giattino, C., & Manheim, D. (2021). Forecasting AI progress: A research agenda. In Technological Forecasting and Social Change (Vol. 170, p. 120909). Elsevier BV. https://doi.org/10.1016/j.techfore.2021.120909

Guillaume, Y. R. F., Dawson, J. F., Otaye-Ebede, L., Woods, S. A., & West, M. A. (2015). Harnessing demographic differences in organizations: What moderates the effects of workplace diversity? In Journal of Organizational Behavior (Vol. 38, Issue 2, pp. 276–303). Wiley. https://doi.org/10.1002/job.2040

Pescetelli, N., Rutherford, A., & Rahwan, I. (2021). Modularity and composite diversity affect the collective gathering of information online. In Nature Communications (Vol. 12, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-021-23424-1

Kapoor, R., & Wilde, D. (2022). Peering into a crystal ball: Forecasting behavior and industry foresight. In Strategic Management Journal. Wiley. https://doi.org/10.1002/smj.3450

Krumrei-Mancuso, E. J., Haggard, M. C., LaBouff, J. P., & Rowatt, W. C. (2020). Links between intellectual humility and acquiring knowledge. The Journal of Positive Psychology, 15(2), 155–170. https://doi.org/10.1080/17439760.2019.1579359

Krumrei-Mancuso, E. J., & Newman, B. (2020). Intellectual humility in the sociopolitical domain. Self and Identity, 19(8), 989–1016. https://doi.org/10.1080/15298868.2020.1714711

Krumrei-Mancuso, E. J., & Rouse, S. V. (2015). The Development and Validation of the Comprehensive Intellectual Humility Scale. In Journal of Personality Assessment (Vol. 98, Issue 2, pp. 209–221). Informa UK Limited. https://doi.org/10.1080/00223891.2015.1068174

Lavrakas, P. (2008). Encyclopedia of Survey Research Methods. Sage Publications, Inc. https://doi.org/10.4135/9781412963947

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and Interpersonal Features of Intellectual Humility. Personality and Social Psychology Bulletin, 014616721769769. https://doi.org/10.1177/0146167217697695

Leary, M. (2018). The psychology of intellectual humility. John Templeton Foundation, 3.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. In Perspectives on Psychological Science (Vol. 10, Issue 3, pp. 267–281). SAGE Publications. https://doi.org/10.1177/1745691615577794

Mellers, B., Tetlock, P. E., & Arkes, H. R. (2019). Forecasting tournaments, epistemic humility and attitude depolarization. Cognition, 188, 19–26. https://doi.org/10.1016/j.cognition.2018.10.021

Porter, T., & Schumann, K. (2018). Intellectual humility and openness to the opposing view. Self and Identity, 17(2), 139–162. https://doi.org/10.1080/15298868.2017.1361861

Tetlock, P. E., & Gardner, D. (2015). Superforecasting: The art and science of prediction. Crown Publishers/Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. In Current Directions in Psychological Science (Vol. 23, Issue 4, pp. 290–295). SAGE Publications. https://doi.org/10.1177/0963721414534257

van Knippenberg, D., & Mell, J. N. (2016). Past, present, and potential future of team diversity research: From compositional diversity to emergent diversity. In Organizational Behavior and Human Decision Processes (Vol. 136, pp. 135–145). Elsevier BV. https://doi.org/10.1016/j.obhdp.2016.05.007

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual Humility: Owning Our Limitations. Philosophy and Phenomenological Research, 94(3), 509–539. https://doi.org/10.1111/phpr.12228

Williams, L. V., & Reade, J. J. (2015). Forecasting Elections. In Journal of Forecasting (Vol. 35, Issue 4, pp. 308–328). Wiley. https://doi.org/10.1002/for.2377

Wunderlich, F., & Memmert, D. (2020). Forecasting the outcomes of sports events: A review. In European Journal of Sport Science (Vol. 21, Issue 7, pp. 944–957). Informa UK Limited. https://doi.org/10.1080/17461391.2020.1793002

Table 1

Differences Between Similar vs. Diverse Rationales Across Four Domains

Domain	Similar		Diverse		Paired-Sample *t*-Test			
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
Climate	.526	.334	.299	.268	5.88	148	< .001	.482
Economy	.493	.298	.313	.232	6.15	148	< .001	.504
Peace/War	.517	.304	.289	.219	7.95	148	< .001	.651

Table 2

Predictions by Domain and Time Horizon

Domain	Individual Stage	Exposure to Similar vs. Diverse Perspectives		Outcome
		Team Stage - Similar	Team Stage - Diverse	
		1-Year Forecasts		
Climate	*M* = 422, *SD* = 3.52 (*n* = 474)	*M* = 422, *SD* = 3.62 (*n* = 227)	*M* = 422, *SD* = 3.4 (*n* = 247)	418.22
Economy	*M* = 12.3, *SD* = 1.25 (*n* = 494)	*M* = 12.1, *SD* = 1.28 (*n* = 241)	*M* = 12.3, *SD* = 1.27 (*n* = 253)	11.6
Peace/War	*M* = 79.0, *SD* = 2.57 (*n* = 439)	*M* = 78.7, *SD* = 2.41 (*n* = 216)	*M* = 79.2, *SD* = 2.69 (*n* = 223)	83
		2-Year Forecasts		
Climate	*M* = 427, *SD* = 6.52 (*n* = 478)	*M* = 427, *SD* = 5.82 (*n* = 233)	*M* = 427, *SD* = 7.01 (*n* = 245)	-
Economy	*M* = 13.2, *SD* = 2.11 (*n* = 498)	*M* = 13.1, *SD* = 2.04 (*n* = 244)	*M* = 13.2, *SD* = 2.14 (*n* = 254)	-
Peace/War	*M* = 81.4, *SD* = 5.57 (*n* = 474)	*M* = 81.7, *SD* = 5.19 (*n* = 236)	*M* = 81.3, *SD* = 5.82 (*n* = 238)	-
		20-Year Forecasts		
Climate	*M* = 453, *SD* = 61.4 (*n* = 479)	*M* = 451, *SD* = 58.3 (*n* = 230)	*M* = 455, *SD* = 56.1 (*n* = 249)	-
Economy	*M* = 13.2, *SD* = 4.73 (*n* = 495)	*M* = 12.9, *SD* = 4.62 (*n* = 243)	*M* = 13.5, *SD* = 4.70 (*n* = 252)	-
Peace/War	*M* = 85.1, *SD* = 24.5 (*n* = 492)	*M* = 88.2, *SD* = 23.8 (*n* = 241)	*M* = 85.7, *SD* = 24.7 (*n* = 251)	-

Table 3

Estimated Ranges by Domain and Time Horizon

Domain	Individual Stage	Exposure to Similar vs. Diverse Perspectives	
		Team Stage - Similar	Team Stage - Diverse
		1-Year Forecasts	
Climate	*M* = 6.76, *SD* = 5.11 (*n* = 474)	*M* = 6.58, *SD* = 4.79 (*n* = 227)	*M* = 7.22, *SD* = 5.49 (*n* = 247)
Economy	*M* = 2.96, *SD* = 2.24 (*n* = 494)	*M* = 2.94, *SD* = 2.31 (*n* = 241)	*M* = 3.03, *SD* = 2.14 (*n* = 253)
Peace/War	*M* = 14.9, *SD* = 9.35 (*n* = 439)	*M* = 14.7, *SD* = 8.84 (*n* = 216)	*M* = 15.3, *SD* = 9.67 (*n* = 223)
		2-Year Forecasts	
Climate	*M* = 9.00, *SD* = 7.70 (*n* = 478)	*M* = 9.11, *SD* = 7.73 (*n* = 233)	*M* = 9.29, *SD* = 7.91 (*n* = 245)
Economy	*M* = 3.45, *SD* = 2.72 (*n* = 498)	*M* = 3.50, *SD* = 2.74 (*n* = 244)	*M* = 3.45, *SD* = 2.60 (*n* = 254)
Peace/War	*M* = 16.2, *SD* = 10.2 (*n* = 274)	*M* = 15.7, *SD* = 9.45 (*n* = 236)	*M* = 16.5, *SD* = 10.2 (*n* = 238)
		20-Year Forecasts	
Climate	*M* = 39.3, *SD* = 43.8 (*n* = 479)	*M* = 38.4, *SD* = 41.6 (*n* = 230)	*M* = 43.4, *SD* = 45.4 (*n* = 249)
Economy	*M* = 5.63, *SD* = 4.53 (*n* = 495)	*M* = 5.61, *SD* = 4.17 (*n* = 243)	*M* = 5.89, *SD* = 4.76 (*n* = 252)
Peace/War	*M* = 27.2, *SD* = 23.2 (*n* = 492)	*M* = 28.7, *SD* = 23.9 (*n* = 241)	*M* = 28.7, *SD* = 23.1 (*n* = 251)

Table 4

Judgmental Accuracy of 1-Year Predictions by Domain

Domain	Individual Stage	Exposure to Similar vs. Diverse Perspectives	
		Team Stage - Similar	Team Stage - Diverse
Climate	M = 4.12, SD = 3.04 (n = 474)	M = 4.21, SD = 3.15 (n = 227)	M = 4.06, SD = 2.93 (n = 247)
Economy	M = 1.01, SD = 0.99 (n = 494)	M = 0.99, SD = 0.97 (n = 241)	M = 1.02, SD = 1.00 (n = 253)
Peace/War	M = 4.27, SD = 2.15 (n = 439)	M = 4.41, SD = 2.14 (n = 216)	M = 4.15, SD = 2.15 (n = 223)

Table 5

Proportion of Belief Updating by Domain

Domain	Exposure to Similar vs. Diverse Perspectives	
	Team Stage - Similar	Team Stage - Diverse
Climate	7.25% (50/690)	5.67% (42/741)
Economy	11.68% (85/728)	6.46% (49/759)
Peace/War	10.10% (70/693)	7.87% (56/712)
Total	9.71% (205/2,111)	6.65% (147/2,212)

Table 6

Feeling Thermometer Scores Toward Liberals, Moderates, and Conservatives

Group	Exposure to Similar vs. Diverse Perspectives	
	Post-Tournament - Similar (n = 252)	Post-Tournament - Diverse (n = 266)
Liberals	M = 53.4, SD = 30.9	M = 49.6, SD = 31.1
Moderates	M = 60.1, SD = 19.8	M = 57.0, SD = 20.3
Conservatives	M = 46.3, SD = 32.0	M = 40.4, SD = 31.3

Note. Feeling thermometer scores range from 0 (“Dislike a great deal”) to 100 (“Like a great deal”).

Figure 1

Example of the Forecasting Tournament User Interface

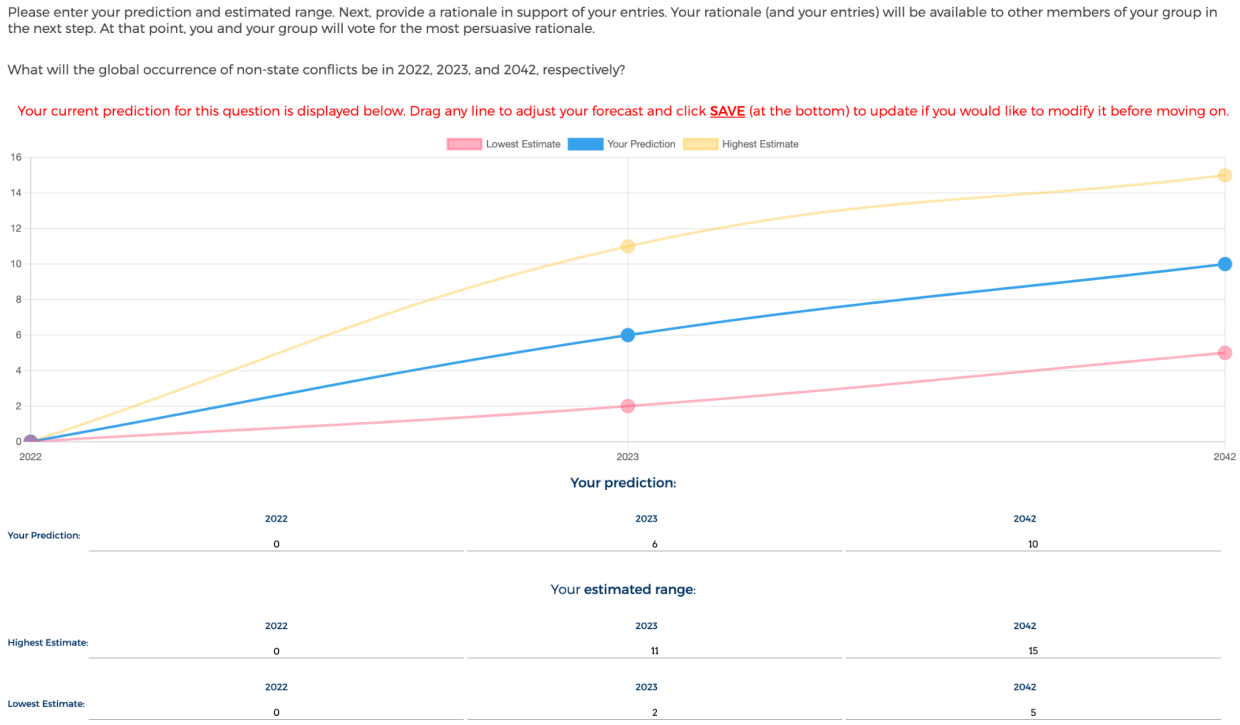


Figure 2

Predictions in the Climate Domain

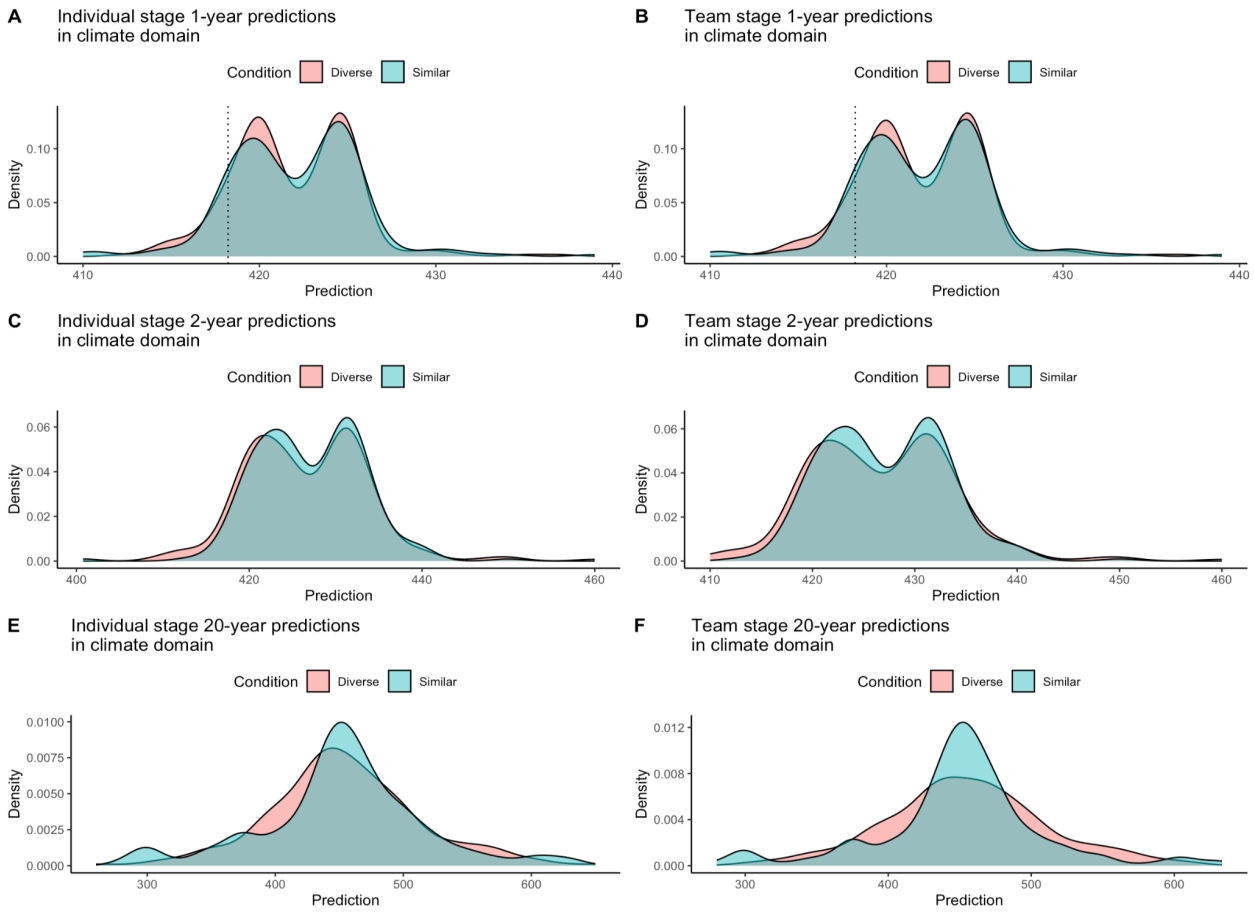


Figure 3

Predictions in the Economy Domain

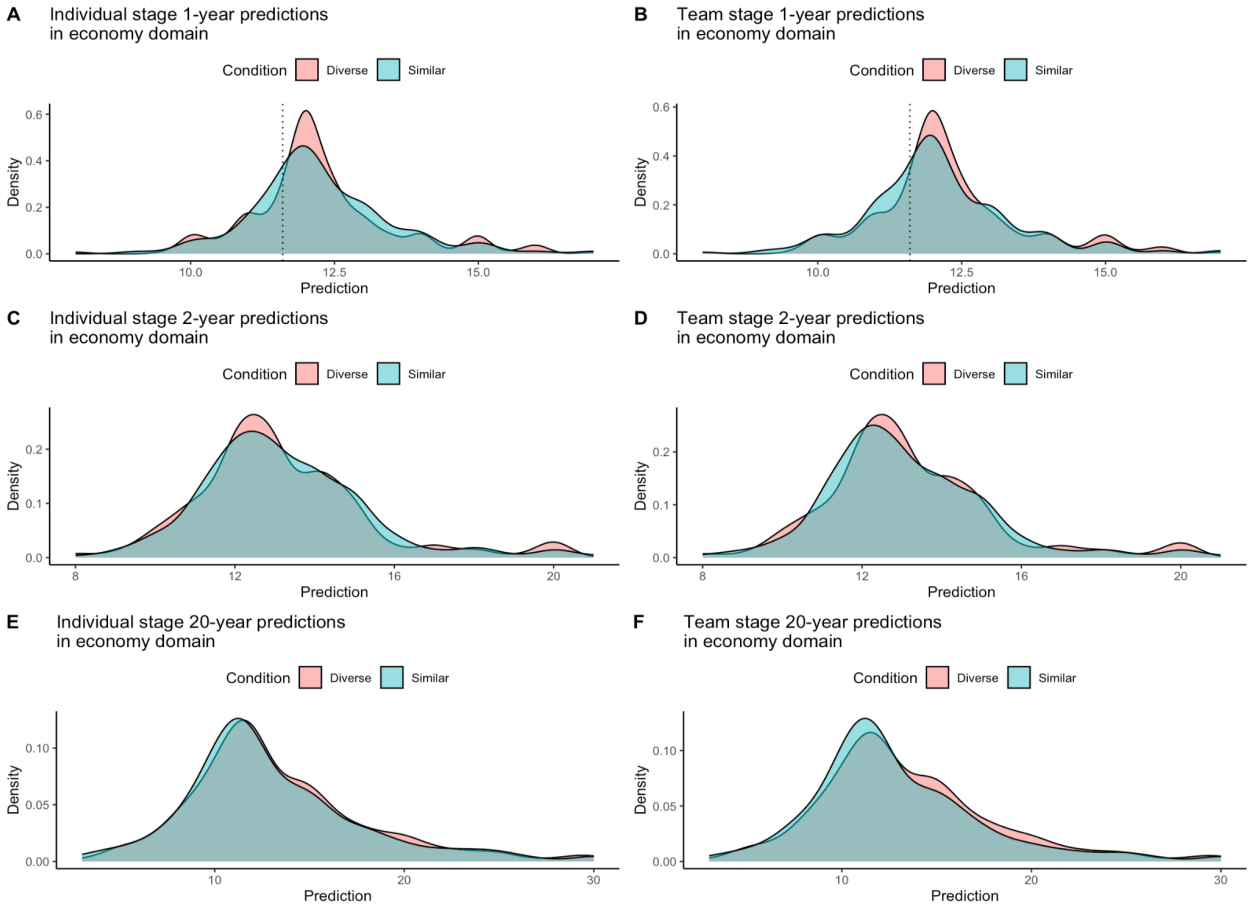


Figure 4

Predictions in the Peace/War Domain

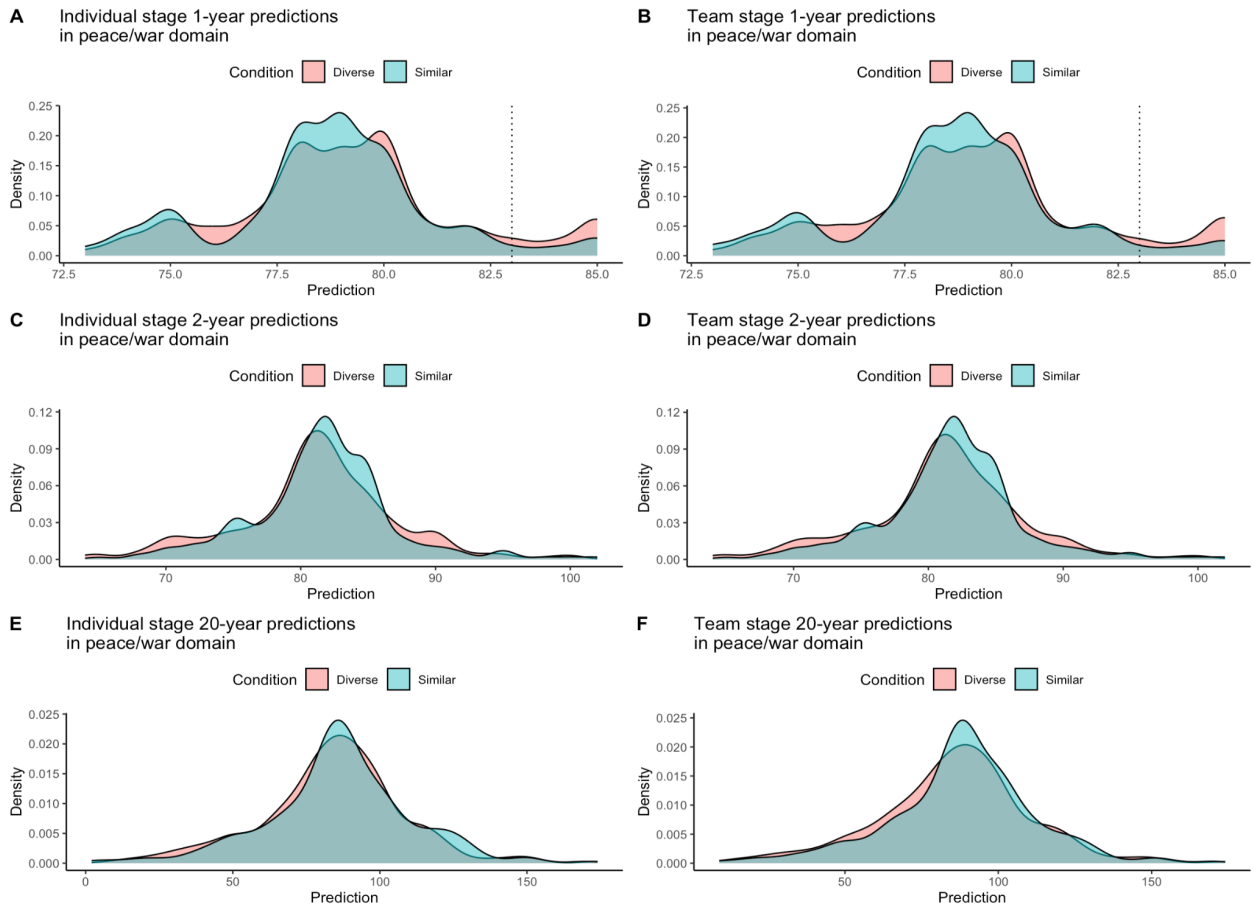


Figure 5

Estimated Ranges in the Climate Domain

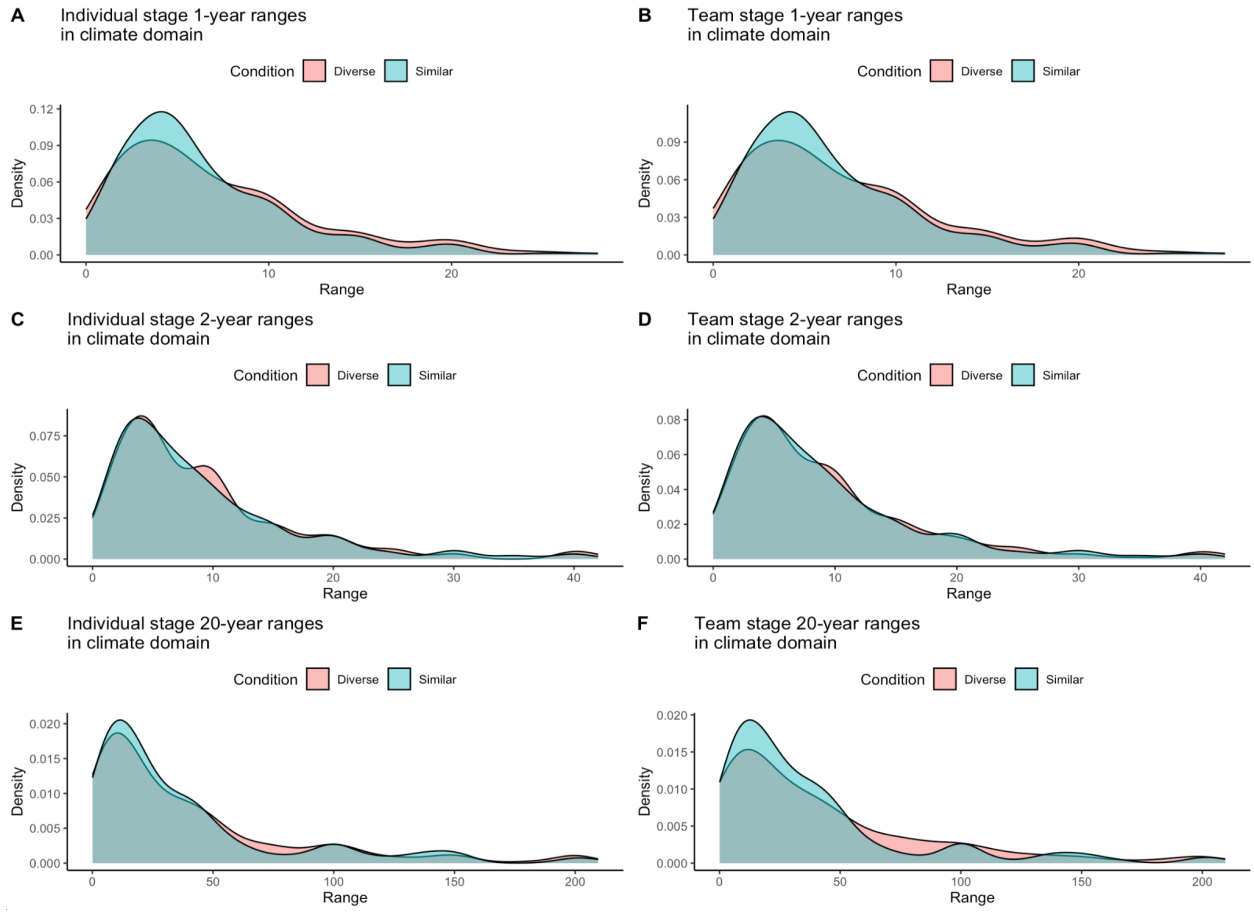


Figure 6

Estimate Ranges in the Economy Domain

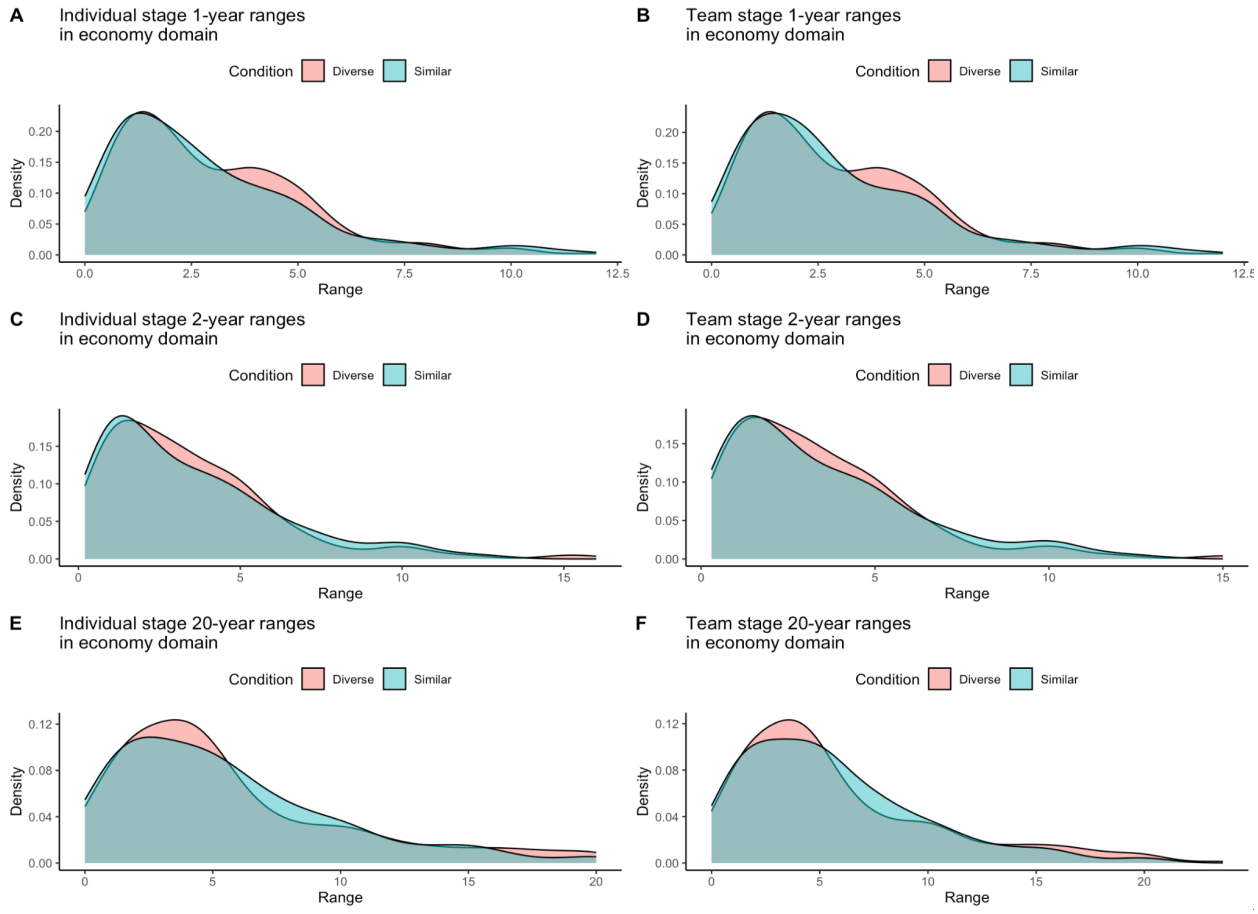


Figure 7

Estimated Ranges in the Peace/War Domain

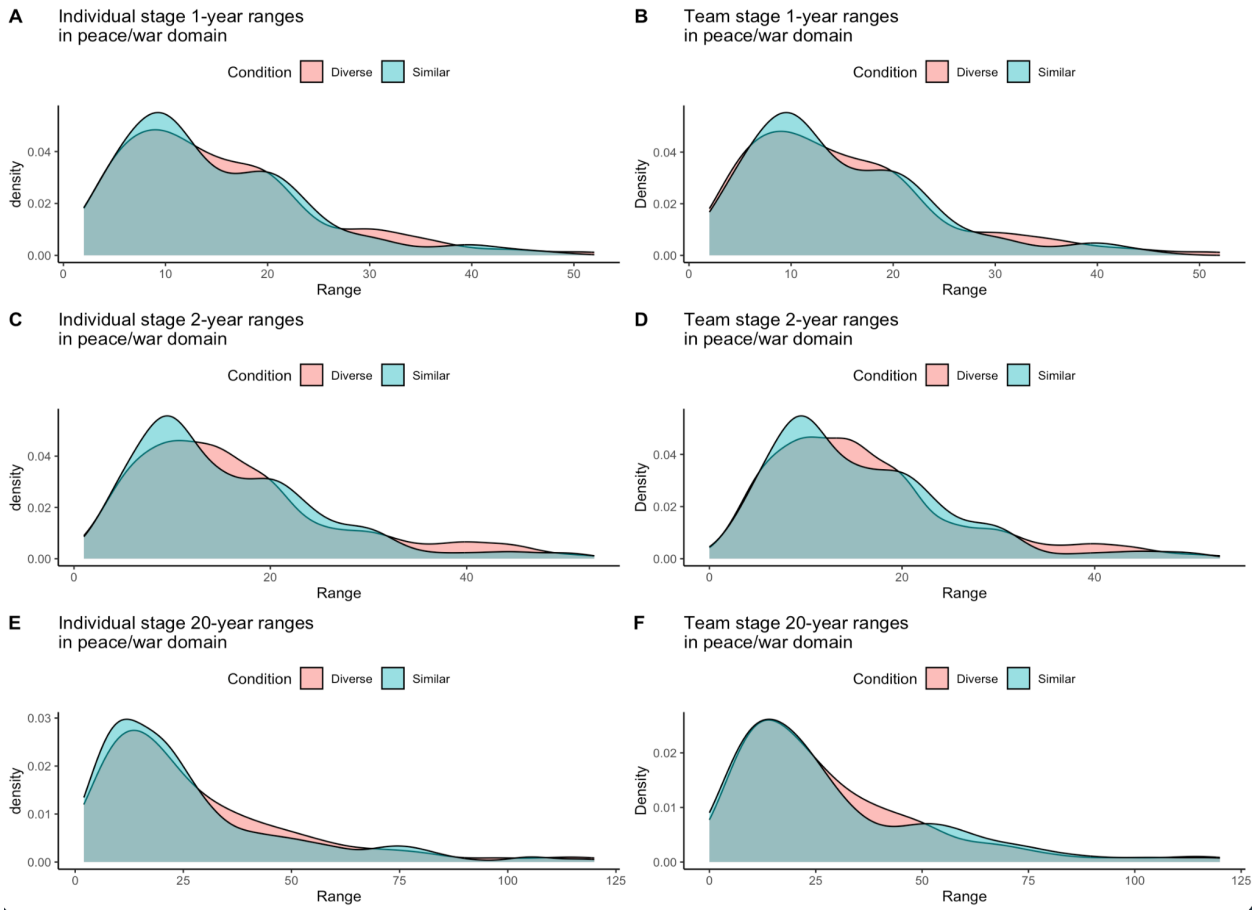
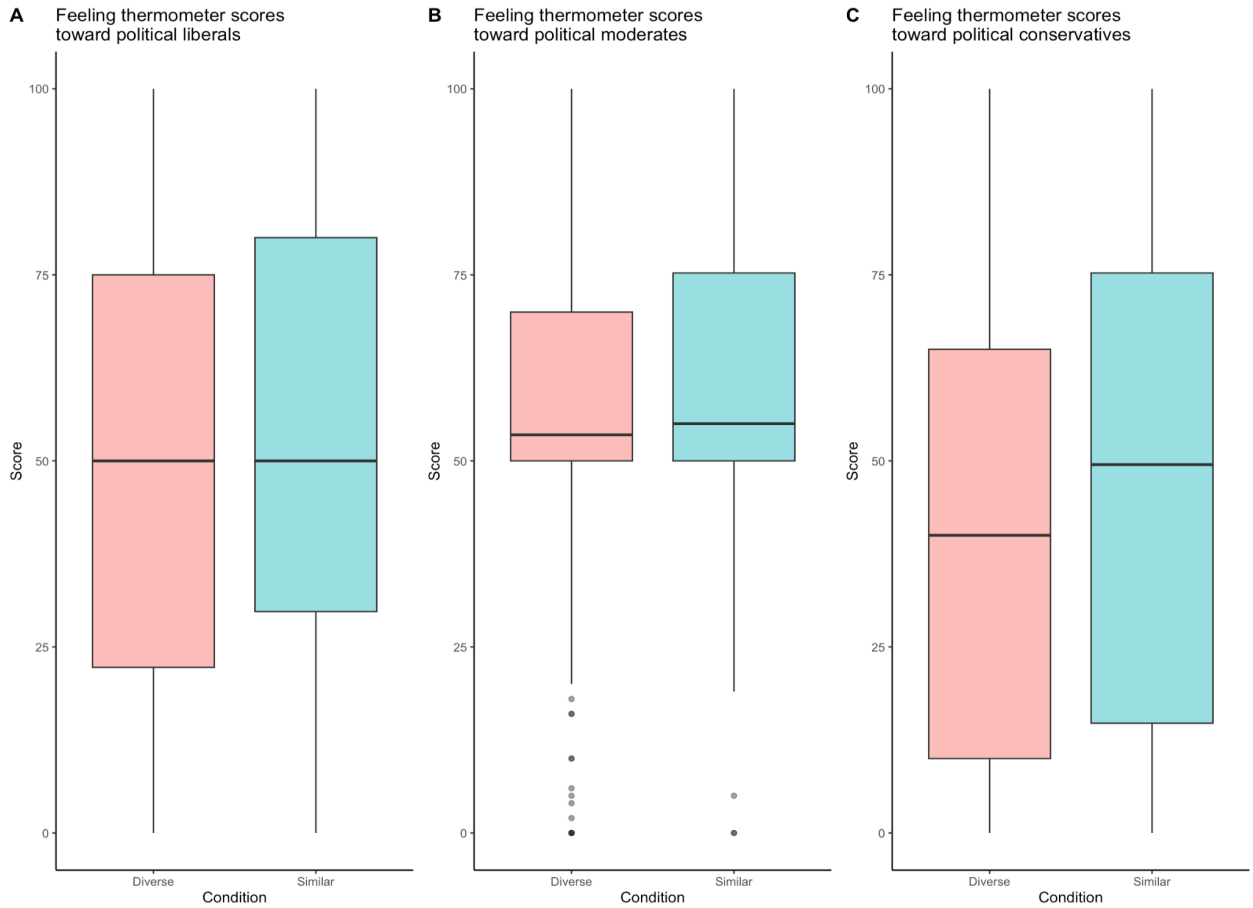


Figure 8

Box Plots of Feeling Thermometer Scores Toward Liberals, Moderates, and Conservatives



Note. Feeling thermometer scores range from 0 (“Dislike a great deal”) to 100 (“Like a great deal”).