Scientific forecasts and reasoning about the COVID-19 pandemic’s societal consequences

Igor Grossmanna, Michael E. W. Varnumb, Cendri Hutchersonc, Amanda Rotellaa, Lyle Ungard, Louis Taye, Philip E. Tetlockd

a – University of Waterloo, Canada

b – Arizona State University, U.S.A.

c – University of Toronto, Canada

d – University of Pennsylvania, U.S.A.

e – Purdue University, U.S.A.

Author note:

Correspondence and material requests should be addressed to Igor Grossmann, [igrossma@uwaterloo.ca](mailto:igrossma@uwaterloo.ca). Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada

# Abstract

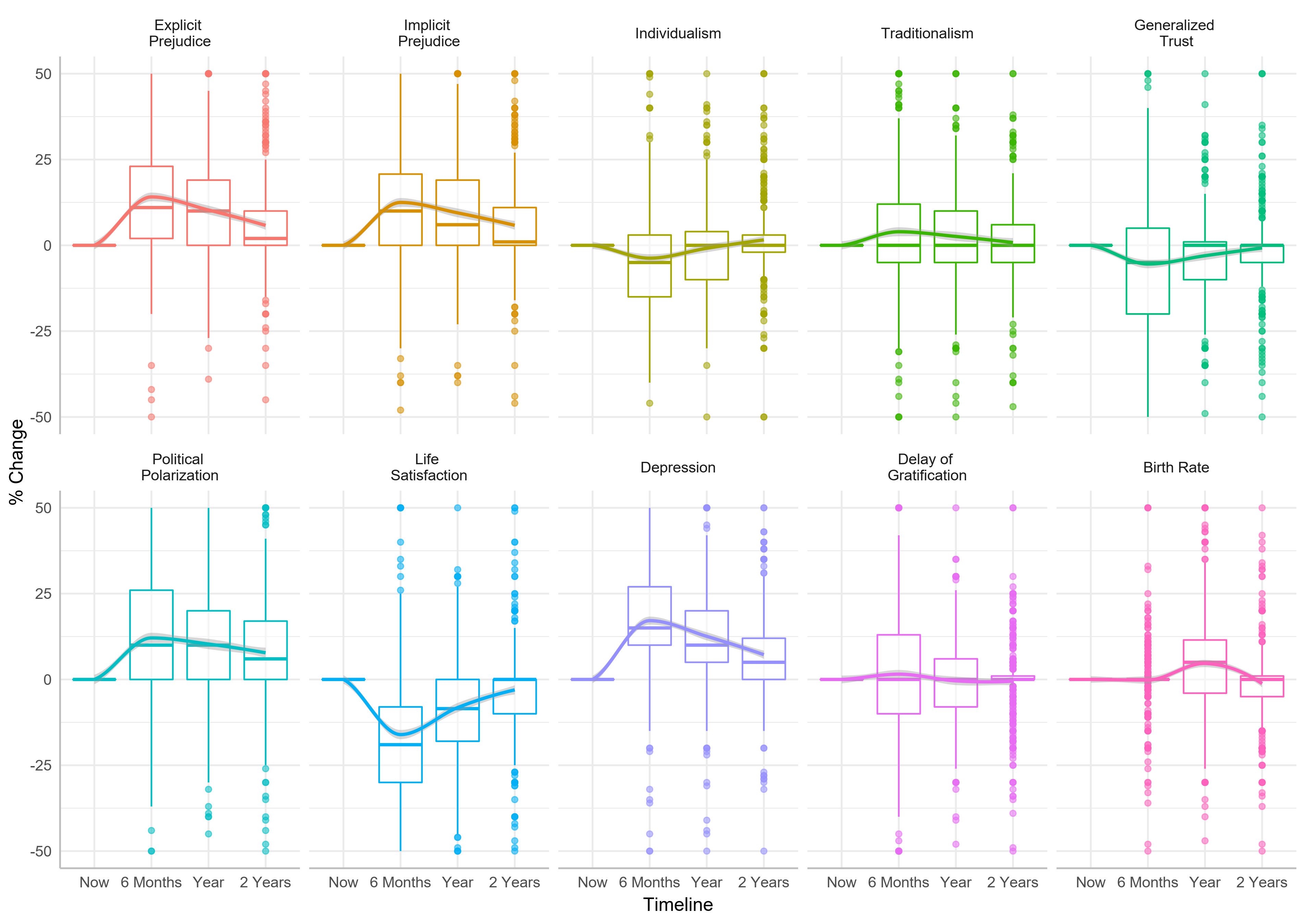
Making predictions is one of the hallmarks of science, but the extent to which behavioral and social scientists can accurately predict societal phenomena and what processes guide their reasoning is unknown. Using the COVID-19 pandemic as a naturalistic experiment, we will explore processes guiding scientists’ ex ante forecasts for phenomena of broad societal relevance, which have been theoretically linked to pathogen threat: gender bias, ideological preferences, political polarization, prejudice, and well-being. In a forecasting tournament,

behavioral and social scientists will receive standardized past data and will submit pre-registered monthly forecasts for a year after the initial peak of the pandemic in the US, with an opportunity to update forecasts based on new data six months later. We will examine forecasting accuracy against benchmarks (e.g., naïve auto-regressive models), and will explore how theoretical, data-driven and exogenous considerations (e.g., expected COVID-19 trajectory) affect scientists’ prediction-oriented reasoning about development of societal phenomena.

Keywords: forecasting; COVID-19; well-being; political polarization; prejudice; metascience

Behavioral and social scientists typically aim to *explain* societal phenomena. At the same time, the utility of scientific theories also involves their ability to accurately *forecast* the future development of a given phenomenon 1,2. A common view across empirical sciences is that theoretical explanations ought to be falsifiable, requiring testing of predictions generated by such theories with future data (or hold-out samples). Historically, the social and behavioral sciences have emphasized the value of explanation rather than prediction 1,3. And in recent years a great deal of effort has been put into testing the extent to which explanations for commonly studied social phenomena are reliable and robust by aiming to estimate the replicability of findings in psychology, economics, and the cognitive sciences 4–8. Yet to date there has been no systematic attempt to assess the extent to which social and behavioral scientists are able to accurately *ex ante* predict -- that is, to forecast -- future trends in the phenomena that they study. We argue that the evaluation of forecasting accuracy can serve as a litmus tests for existing theories and data-modelling approaches aimed at understanding societal dynamics and can provide useful ways to assess their utility in the context of policy recommendations. Moreover, justifications used in the process of making predictions can shed light on scientists’ reasoning about their theoretical models.

Here, we take advantage of a kind of natural experiment, the outbreak of the COVID-19 pandemic, to conduct a standardized evaluation of forecasting accuracy 1 among behavioral and social scientists in well-studied domains which are believed to be linked to the threat of infectious disease and for which rich longitudinal data is available: gender bias, ideological preferences, outgroup prejudice, political polarization, and well-being. Further, whereas some economic and behavioral consequences of the COVID-19 pandemic appear straightforward, forecasts about societal consequences for these domains have compelling alternatives. For instance, the pandemic could lead to societal decline in happiness and mental health 9, but it could also lead to resilience in the face of adversity 10,11 or an enhanced sense of meaning 12. Threat of an infectious disease can promote xenophobia, prejudice, and changes in political preferences 13–16, but it could also bring different groups together given the external nature of the threat 17. Speculations about social consequences of the pandemic are not confined to journalists and political pundits but are also being made by social and behavioral scientists. Recently we made an initial attempt to quantify these intuitions in a survey of over 300 behavioral and social scientists conducted in early April 2020. We found that predictions concerning possible changes in cultural values, generalized trust, political polarization, prejudice or well-being varied greatly, with some experts predicting up to 50% change in outgroup prejudice, political polarization, or depression (see Fig. 1).



**Fig. 1**. Preliminary survey of behavioral and social scientists (*N* = 386) concerning their predictions for societal changes in the US over the course of the COVID-19 pandemic. Boxplots indicating distribution of predictions for each time point. Loess line of best fit and 95% confidence interval show estimated change over time. Predictions appear to vary widely within and across domains: Whereas predictions for well-being, prejudice and political ideology suggest substantial (> 10%) shifts in the next year, predictions for social shifts in values, trust, or self-regulation appear modest. For further details, see <https://osf.io/vqd4a/>

How accurate are the forecasts of social and behavioral scientists in general? Past research on behavioral, geopolitical, and public policy forecasts paints a complex picture. On the one hand, classic work suggests that experts under many circumstances perform worse than statistical, data-driven forecasts 18. This is noteworthy, given recent evidence that social scientists’ data-driven predictions for individuals’ life outcomes appear largely inaccurate 19. In a similar vein, individual public policy experts also appear to provide poorer forecasts than do expert groups 20. On the other hand, expertise in social and behavioral issues appears to improve forecasting accuracy for geopolitical events 21 and behavior 22. Notably, prior forecasting work has chiefly focused on predictions for specific geopolitical events, economic trends, and individual-level behaviors. Furthermore, whereas some recent work has begun to examine prediction accuracy of supervised machine-learning models for individual life outcomes against a hold-out sample 19, such work has not examined predictions ex ante. In fact, to our knowledge no work has systematically compared the accuracy of ex ante forecasts derived via machine-learning methods to forecasts based on theory (or a combination of theory and modeling). Moreover, relatively little is known about how behavioral and social scientists go about making such forecasts, including how they use prior theory, and how they update their theoretical models and predictions in light of new data.

Inspired by prior crowdsourced tournaments among researchers 19,23 and prior forecasting research 20,24, we will conduct a crowdsourced forecasting tournament under the framework of the Forecasting Collaborative to provide an initial estimate of forecasting accuracy in social and behavioral sciences and unpack the reasoning processes that guide such scientists when they engage in forecasting. Although the present tournament takes place during a pandemic and is focused on outcomes that have in some way been previously linked to levels of infectious disease, this tournament may nonetheless provide some general insight into these questions and will set the stage for further efforts in the future.

The Forecasting Collaborative is open to all scientists who wish to participate in the tournament and who provide forecasts for the next 12 months for the domains specified. To ensure a “common task framework” 1,19, we will provide all participants with the same time series data for each dependent variable going back 39 months, which they may use to inform their forecasts. Teams will not be constrained in terms of the methods they use to generate timepoint forecasts. However, they will be required to specify their methods and, if they make use of data-driven methods, they will be required to provide the model and any additional data that they use to generate their forecasts.

Subsequently, project leaders will evaluate and compare different types of forecasts to each other and against simple extrapolation algorithms (e.g., moving averages at different lags) as benchmarks to identify which forecasting approaches lead to greater accuracy in each of the forecasted domains, and whether there are any general rules or insights that can be gleaned in terms of forecasts within or across these domains. In other words, we will be able to evaluate whether scientists’ forecasts are better than naïve extrapolation from prior data. In addition, to assess and protect against the possibility that forecasting models are accurate by chance (in the same way that some investing strategies can “get lucky” in a particular year without actually being better than other strategies), we will use subsets of the data to determine whether model accuracy in one subset of predictions correlates with model accuracy in other subsets. For example, we will test whether rankings of models with respect to accuracy in predicting the first six months of values is similar to model rankings when assessing the next six months.

This tournament will be the first of its kind to evaluate the accuracy of social scientists’ ex ante point forecasts for key societal phenomena. Further it will provide insight not only into which kinds of approaches generate more accurate forecasts for future societal dynamics, but also into reasoning processes and justifications guiding scientists’ forecasts.

**Research Questions**

The Forecasting Collaborative tournament will provide an opportunity to evaluate behavioral and social scientists’ abilities to accurately forecast development of major societal issues over the course of the COVID-19 pandemic. Participants will be able to select up to ten social issues for which 39 months (from January 2017 to March 2020) of prior longitudinal monthly data exists: affective well-being and life satisfaction, ideological preferences, political polarization, explicit and implicit attitudes towards Asian-Americans, explicit and implicit attitudes towards African Americans, and explicit and implicit stereotypes concerning gender and career. Data from the tournament will allow systematic comparisons of pre-registered theoretical and data-driven forecasting approaches. Results from the tournament will address the following questions:

1) How good are behavioral and social scientists at forecasting the social consequences of a COVID-19 pandemic? Following established procedures 1, we will examine the absolute percentage deviation for each forecast, and mean absolute scaled error (MASE) 25 within and across forecasted time-points and social issues. MASE compares forecasted values against those obtained via a one-step “naïve forecast method.” It is independent of the scale of the data and can be used to compare forecasts across datasets with different scale, is asymptotically normal and easy to interpret, with lowest MASE scores indicating greatest forecasting accuracy. Critically, we will compare forecasting accuracy of scientists’ predictions against basic interpolation algorithms (e.g., moving average models with different lags). We will also compare the stability of model accuracy measured for different subsets of time, to assess the extent to which models might be accurate simply by chance.

2) Are some societal shifts in response to the COVID-19 pandemic easier to accurately forecast than others (e.g., is it easier to accurately forecast changes in prejudice toward outgroups vs. well-being vs. shifts in political preferences)? We will examine overall forecasting accuracy, and stability of forecasting accuracy, across domains above and beyond the naïve forecasting method using MASE 25.

3) Are there characteristics (discipline, methodological approach to forecasts) of some teams that lead to more accurate forecasts in social science domains? Here, we focus on comparisons of forecasting approaches relying on a. pure expertise (but no data modelling); b. pure modelling (but no consideration of expert theories); c. hybrid approaches. We will explicitly examine the reasoning process, evaluating the role of confidence, and quality of forecast rationales (e.g., consideration of conditional factors and counterfactuals) for forecasting accuracy. Examination of additional variables included in the models, such as the trajectory of the COVID-19 pandemic as well as other exogenous factors specified by the teams, will allow us to assess whether and how much consideration of key exogenous variables may enhance forecasting accuracy. This focus on scientists’ reasoning about their forecasts will also enable us to address questions about whether forecasts may be wrong for the right reasons. Finally, we will also consider social factors such as the team composition (*N*, disciplinary diversity). We will compare the accuracy of forecasts generated via these different approaches by comparing groups’ absolute forecasting deviations, as well as their MASE scores and margin of error within each class of forecasts. These sorts of analyses are essential to understanding at a process level how scientific knowledge advances but have not previously been systematically addressed in other work on forecasting, especially among scientists.

Within the tournament, we will also deploy and evaluate the effectiveness of various strategies for improving forecasting accuracy. For example, we will provide teams with the opportunity to compare their forecasts to new data at a later timepoint and to update their predictions 24. We will also provide teams with a summary of diverse forecasting rationales from other teams at the half-way point of the tournament, giving them additional information which they may use to revise their forecasts. We will compare updated vs. original forecasts to evaluate changes in forecasting accuracy and will ask teams to provide a rationale for these updates. This will allow us to distinguish between Bayesian-style updating (i.e., the influence of new data on posterior predictions) and theory-driven updating (i.e., the influence of shared scientific ideas and knowledge on the development of new predictions).

# Method

## Ethics

The study has been approved by the Office of Research Ethics of the University of Waterloo under protocol # 42142.

**Forecasting Domains and Data Pre-Processing**

Computational forecasting models require enough prior time series data for reliable modelling. Based on prior recommendations 26, we will provide each team with at least 39 monthly estimates for each of the domains of their choice, which can be used for data-driven forecasting and for establishing a baseline estimate prior to the U.S. peak of the pandemic. The monthly estimates will encompass a time window from Jan 2017 to March 2020.

Because of the requirement for rich standardized data for computational approaches to forecasting 1, we limit forecasting domains to issues of broad societal significance for which large-scale longitudinal monthly time series data can be obtained. Such domains include affective well-being and life satisfaction, political ideology and polarization, bias in explicit and implicit attitudes towards Asian-Americans and African-Americans, as well as stereotypes regarding gender and career vs. family. Moreover, there are broad social consequences associated with these issues. Developing accurate models to predict how they unfold over time would be valuable to behavioral and social scientists across fields.

To establish a “common task framework” 1, we standardized methods for obtaining relevant prior data for each of these domains.

**Affective well-being and life satisfaction**. We used monthly Twitter data to estimate markers of affective well-being (positive and negative affect) and life satisfaction over time. We rely on Twitter because no polling data for monthly well-being over the required time period exists, and because prior work suggests that national estimates obtained via social media language can reliably track subjective well-being 27. For each month, we used previously validated predictive models of well-being, as measured by affective well-being and life satisfaction scales 28. Affective well-being was calculated by applying a custom lexicon 29 to message unigrams; life satisfaction was estimated using a ridge regression model trained on *latent Dirichlet allocation* topics, selected using univariate feature selection and dimensionally reduced using randomized principal component analysis, to predict Cantril ladder life satisfaction scores. Such twitter-based estimates closely follow nationally representative polls 30. We applied the respective models to Twitter data from January 2017 to March 2020 to obtain estimates of affective well-being and life satisfaction via language on social media.

**Ideological Preferences***.* We approximated monthly ideological preferences via aggregated weighted data from the Congressional Generic Ballot polls conducted between January 2017 and March 2020 ([projects.fivethirtyeight.com/congress-generic-ballot-polls](https://projects.fivethirtyeight.com/congress-generic-ballot-polls)), which ask representative samples of Americans to indicate which party they would support in an election. We weighted polls based on FiveThirtyEight pollster ratings, poll sample size, and poll frequency. FiveThirtyEight pollster ratings are determined by their historical accuracy in forecasting elections since 1998, participation in professional initiatives that seek to increase disclosure and enforce industry best practices and inclusion of live-caller surveys to cellphones and landlines. Based on this data, we then estimated monthly averages for support of Democrat and Republican parties across pollsters (e.g., Marist College, NBC/Wall Street Journal, CNN, YouGov/Economist).

**Political Polarization**. We assessed political polarization by examining differences in presidential approval ratings by party identification from Gallup polls (<https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx>). We obtained a difference score in % of Republican versus Democrat approval ratings and estimated monthly averages for the time period of interest. The absolute value of the difference score will ensure possible change after 2020 Presidential election will not change the direction of the estimate.

**Explicit and Implicit Bias**. Given the natural history of the COVID-19 pandemic, we sought to examine forecasted *bias in attitudes towards Asian-American* (vs. European-Americans). To further probe racial bias, we sought to examine forecasted *racial bias* in preferences for African-American (versus European-American) people. Finally, we sought to examine *gender bias* in associations of female (vs. male) gender with family versus career. For each domain we sought to obtain both reliable estimates of explicit attitudes 31 and estimates of implicit attitudes 32. To this end, we obtained data from the Project Implicit website (http://implicit.harvard.edu) which has collected continuous data concerning explicit stereotypes and implicit associations from a heterogeneous pool of volunteers (50,000 - 6,000 unique tests on each of these categories per month). Further details about the website and test materials, are publicly available at <https://osf.io/t4bnj>. Recent work suggests that Project Implicit data can provide reliable societal estimates of consequential outcomes 33,34 and when studying cross-temporal societal shifts in U.S. attitudes 35. Despite the non-representative nature of the Project Implicit data, recent analyses suggest that bias scores captured by Project Implicit are highly correlated with nationally representative estimates of explicit bias, *r*  = .75, indicating that group aggregates of the bias data from Project Implicit can reliably approximate group-level estimates 34. To further correct possible non-representativeness, we applied stratified weighting to the estimates, as described below.

For explicit attitude scores, participants provided ratings on feeling thermometers towards Asian-Americans and European Americans (to assess Asian-American bias), and White and Black Americans (to assess racial bias). For explicit bias in the Gender – Career task, participants rated the extent to which they associated career with male or female (from *Strongly Female* to *Strongly Male*) and then used the same scale to rate the extent to which they associated family with male or female. Relative explicit bias was then calculated as the difference in responses to minority and majority groups on feeling thermometers (for Asian-American and racial bias) and the family and career items (for gender bias).

Implicit attitude scores were computed using the revised scoring algorithm of the implicit association test (IAT) 36. The IAT is a computerized task comparing reaction times to categorize paired concepts (in this case, social groups, e.g., Asian American vs. European American) and attributes (in this case, valence categories, e.g., good vs. bad). Average response latencies in correct categorizations were compared across two paired blocks in which participants categorized concepts and attributes with the same response keys. Faster responses in the paired blocks are assumed to reflect a stronger association between those paired concepts and attributes. In all tests, positive IAT *D* scores indicate a relative preference for the typically preferred group.

Respondents whose scores fell outside of the conditions specified in the scoring algorithm did not have a complete IAT *D* score and were therefore excluded from analyses. Restricting the analyses to only complete IAT *D* scores resulted in an average retention of 92% of the complete sessions across tests. The sample was further restricted to include only respondents from the United States to increase shared cultural understanding of attitude categories. The sample was restricted to include only respondents with complete demographic information on age, gender, race/ethnicity, and political ideology.

We used explicit and implicit bias data for January 2017 – March 2020 and created monthly estimates for each of the explicit and implicit bias domains. Because of possible selection bias among the Project Implicit participants, we adjusted population estimates by weighting the monthly scores based on their representativeness of the demographic frequencies in the U.S. population (age, race, gender, education; estimated biannually by the U.S. Census Bureau; <https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html>). Further, we adjusted weights based on political orientation (1 = “strongly conservative;” 2 = “moderately conservative;” 3 = “slightly conservative;” 4 = “neutral;” 5 = “slightly liberal;” 6 = “moderately liberal;” 7 = “strongly liberal”), using corresponding annual estimates from the General Social Survey. With the weighting values for each participant, we computed weighted monthly means for each attitude test. These procedures ensured that weighted monthly averages approximated the demographics in the U.S. population. We cross-validated this procedure by comparing weighted annual scores to nationally representative estimates for feeling thermometer for African-American and Asian-American estimates from the American National Election studies in 2017 and 2018.

For each of these domains, forecasters will be provided with 39 monthly estimates, as well as detailed explanation about the origin and the calculation of respective indices. Thereby, we aim to standardize the data source for the purpose of the forecasting competition 1. See Appendix for examples worksheets provided to participants for submissions of their forecasts.

## Participants

In this work we aim for a minimum sample of 40 researchers/teams in a forecasting tournament after prescreening to ensure that participants possess at minimum a bachelor’s degree in behavioral, social, or computer sciences. Using GPower 3.1, we estimated that for a typical effect size for the forecasted social issues 37, *f* = .14, with 12 measurement points provided by participants (reflecting forecasts for 12 months), with 80% power, and an expected correlation among repeated times series data points of .7, we would need 33 scientist teams for the tournament to statistically compare accuracy among three groups (experts vs. data-based forecasts vs. hybrid-based forecasts).

Participants will be recruited via large scale advertising on social media, mailing lists in the behavioral and social sciences, decision sciences, and data science, advertisement on academic social networks including ResearchGate, and through word of mouth. To ensure broad representation across the academic spectrum of relevant disciplines, we will target groups of scientists working on computational modeling, social psychology, judgment and decision-making, and data science. The Forecasting Collaborative will start by the end of April 2020, during which time the U.S. Institute for Health Metrics and Evaluation projects the peak of the COVID-19 pandemic in the US to occur. The recruitment phase will continue for one month. If 40 teams have not joined the tournament by this time, we will extend the deadline for entry until 40 researchers/teams have joined.

## Procedure

Information for this project is available on the designated website (predictions.uwaterloo.ca), which includes objectives, instructions, and prior monthly data for each of the 10 domains they can use for modelling. If researchers decide to partake in the tournament, they can sign up via an active Qualtrics survey, which will ask them to upload their estimates for forecasting domains of their choice (see Appendix) and answer a set of confidential questions about their rationale and forecasting team composition. Once all data is received, de-identified responses will be used to pre-register the forecasted values and models on the Open Science Framework. At the half-way point (i.e., at six months), participants will be provided with a comparison summary of their initially forecasted point estimates vs. actual data for the initial six months, as well as a summary of diverse rationales for forecasts from other teams. Subsequently, they will be provided with an option to update their forecasts, provide a detailed description of the updates, and answer an identical set of questions about their data model and rationale for their forecasts, as well as the consideration of possible exogenous variables and counterfactuals.

**Forecasting justifications.**  For each forecasting model submitted to the tournament, participants will provide detailed descriptions. They will describe the type of model they computed (e.g., time series, game theoretic models, other algorithms), model parameters, additional variables they included in their predictions (e.g.., COVID-19 trajectory, presidential election outcome), and underlying assumptions. Additional parameters can be continuous variables (e.g., COVID-19 deaths; unemployment rate) or based on a single discrete event (e.g., political leadership change; implementation of a policy measure). Participants will also provide a theoretical justification for these decisions.

**Confidence.** Participants will rate their confidence in their forecasted points for each forecast model they submit. Confidence will be rated on a 7-point scale from 1 (not at all) to 7 (extremely).

**COVID-19 Conditional.** Next, we will zero-in on the COVID-19 pandemic as a conditional of interest given links between infectious disease and the target social issues we selected for this tournament. Continuous real-time data for this variable is being currently being gathered and will continue to be available over the course of the forecasting tournament. Participants will report if they used the past or predicted trajectory of the COVID-19 pandemic (as measured by number of deaths or prevalence of cases or new infections) as a conditional in their model, and if so will provide their forecasted estimates for the COVID-19 variable included in their model.

**Counterfactuals.** Counterfactuals are hypothetical alternative historic events that would be thought to affect the forecast outcomes, if they were to occur. Participants will describe the key counterfactual events between December 2019 and April 2020 that they theorize would have led to different forecasts (e.g., U.S.-wide implementation of social distancing practices in February). Two independent coders will evaluate the distinctiveness of counterfactuals. If discrepancies arise, they will discuss individual cases with other members of the forecasting collaborative to make the final evaluation.

**Team characteristics**. To assess objective expertise, teams will report if any of their members have previously researched or published on the topic of their forecasted variable. They will also report each member's areas of expertise and amount of education. To assess subjective expertise, teams will report their agreement with the statement: “My team has strong expertise on the research topic of Life Satisfaction.”

**Data Analysis Plan**

**Categorization of Forecasts**

We will categorize forecasts based on modeling approaches. Specifically, two independent research associates will categorize forecasts for each domain based on provided justifications: i. purely based on (a) theoretical model(s); ii. purely based on data-driven model(s); iii. a combination of theoretical and data-driven models – i.e., computational model relies on specific theoretical assumptions. We will further identify modelling approaches that solely rely on extrapolation of time series from the data we provided (e.g., ARIMA, moving average with lags; yes/no). Disagreements between coders here and below will be resolved through joint discussion with the leading three authors of the project.

**Categorization of Additional variables**

We will test how the presence and number of additional variables as parameters in the model impact forecasting accuracy. To this end, we will ensure that additional variables are distinct from one another. Two independent coders will evaluate the distinctiveness of each reported parameter. When there are discrepancies arise, the coders will discuss the case with lead members of the forecasting collaborative to arrive at a consensus.

## Categorization of Teams

We will next categorize teams based on compositions. First, we will sort contributors into three categories: i. singular forecaster; ii. small group (*n* < 6); iii. large group (*n* ≥ 6). Next, we will sort teams based on disciplinary orientation: i. behavioral sciences; ii. social sciences; iii. computer sciences; iv. interdisciplinary/other. Finally, we will use information teams provided concerning their objective and subjective expertise level for a given subject domain. We will use each covariate in separate multi-level analyses with domains and time points as predictors and absolute percentage error scores for a given forecast as a dependent variable.

**Forecasting Update Justifications**

Given that participants will receive both new data and a summary of diverse theoretical positions they can use as a basis for their updates, two independent research associates will score participants’ justifications for forecasting updates on three dummy-categories: i. new six months of data we provide; ii. theoretical insights from the summary of teams’ rationales we provide; iii. consideration of other external events.

## Confirmatory Analyses: Comparison of Forecasting Models

We will first investigate overall forecasting accuracy in behavioral and social sciences by examining MASE for each of the forecasting domains. Using MASE scores will allow us to compare forecasted models against the naïve baseline model.

**Exploratory Analyses: Comparison of Different Approaches/Teams**

The main exploratory (two-tailed) analyses will compare MASE scores for the whole forecasted time series as well as percent of absolute error for each individual forecasted time point when using different forecasting approaches. To this end, we will fit a series of linear mixed effect models. For models evaluating overall accuracy of the forecasted model, we will use forecasting type (purely theoretical, purely data-driven and hybrid models), forecasting domain as predictors, with MASE scores nested within teams. Next, we will examine how the theory-free “extrapolation of time series” models compare in forecasting accuracy to models that rely on other model parameters and/or theoretical assumptions, by including this contrast between models and forecasting domain as predictors, with MASE scores nested within teams. For models evaluating accuracy of individual time points, we will use forecasting type (purely theoretical, purely data-driven and hybrid models), forecasting domain and time points as predictors, with absolute percent deviation scores nested within teams.

We will use equivalent analyses with team type and confidence (instead of forecasting type) as predictors. Further, we will examine whether presence of additional parameters (beyond time series data we provide) and counterfactuals significantly alters forecasting accuracy. First, in a series of linear mixed models similar to the one outlined above we will examine whether presence (dummy-coded yes/no) or number of considered additional parameters and counterfactuals moderate the forecasting accuracy (MASE scores for total accuracy / percent of absolute error for accuracy at specific time points).

Next, we will zero-in forecasts including COVID-19 virus trajectory as a conditional. For these forecasts, we will first estimate the forecasting accuracy of the COVID-19 trajectory by evaluating MASE scores for COVID-19 death against the actual number of deaths. We will use these conditional forecasting accuracy scores as a moderator in linear models evaluating accuracy of each of the targeted domain. We will further conduct simple slope analyses, evaluating the role of conditional forecasting accuracy for the accuracy of the forecast in targeted domains. Such analyses can reveal whether participants’ forecasting errors in targeted domains may be qualified by their accuracy in expectations for the virus trajectory.

For categorical predictors we will perform post-hoc Scheffe pair-wise comparisons to identify significant differences in forecasting accuracy across categories. For all analyses, we will assess the robustness of our conclusions by examining forecasting accuracy not only when using the whole timeseries, but also when using different subsets of time (e.g. first six months, last six months).

**Evaluating Updating-contingent Change in Forecasting Accuracy**

We will examine whether teams become more accurate on average if they decide to update their forecasts and will also evaluate predictors of improved forecasting accuracy.

We will fit a series of linear mixed effect models. To evaluate improved accuracy of the overall forecast, we will use linear mixed effects models with group type (update vs. no update) and forecasting domain as predictors and MASE scores for updated forecasts for the remaining six months as a dependent variable, nested within teams. We will follow-up with post-hoc contrast tests, in which we will compare forecasting accuracy of different updating strategies (i. new six months of data we provide; ii. theoretical insights from the summary of teams’ rationales we provide; iii. consideration of other external events) against each other and against the no-updating control group.

# References

1. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science (80-. ).* **355**, 486–488 (2017).

2. Varnum, M. E. W. & Grossmann, I. Cultural Change: The How and the Why. *Perspect. Psychol. Sci.* **12**, 956–972 (2017).

3. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology : Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).

4. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science (80-. ).* **349**, aac4716–aac4716 (2015).

5. Klein, R. A. *et al.* Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

6. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science (80-. ).* **351**, 1433–1436 (2016).

7. Cova, F. *et al.* Estimating the Reproducibility of Experimental Philosophy. *Rev. Philos. Psychol.* (2018). doi:10.1007/s13164-018-0400-9

8. Schauer, J. M. & Hedges, L. V. Assessing heterogeneity and power in replications of psychological experiments. *Psychol. Bull.* (2020). doi:10.1037/bul0000232

9. Diener, E., Lucas, R. E. & Scollon, C. N. Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *Am. Psychol.* **61**, 305–314 (2006).

10. Anusic, I. & Schimmack, U. Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *J. Pers. Soc. Psychol.* **110**, 766–781 (2016).

11. Bonanno, G. A. Clarifying and Extending the Construct of Adult Resilience. *Am. Psychol.* **60**, 265–267 (2005).

12. Jayawickreme, E. & Blackie, L. E. R. Post-traumatic Growth as Positive Personality Change: Evidence, Controversies and Future Directions. *Eur. J. Pers.* **28**, 312–331 (2014).

13. Fincher, C. L. & Thornhill, R. Parasite-stress promotes in-group assortative sociality: The cases of strong family ties and heightened religiosity. *Behav. Brain Sci.* **35**, 61–79 (2012).

14. Varnum, M. E. W. & Grossmann, I. Pathogen prevalence is associated with cultural changes in gender equality. *Nat. Hum. Behav.* **1**, 0003 (2016).

15. Schaller, M. & Murray, D. R. Pathogens, personality, and culture: Disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *J. Pers. Soc. Psychol.* **95**, 212–221 (2008).

16. van Leeuwen, F., Park, J. H., Koenig, B. L. & Graham, J. Regional variation in pathogen prevalence predicts endorsement of group-focused moral concerns. *Evol. Hum. Behav.* **33**, 429–437 (2012).

17. Niou, E. M. S. & Tan, G. External threat and collective action. *Econ. Inq.* **43**, 519–530 (2005).

18. Dawes, R., Faust, D. & Meehl, P. Clinical versus actuarial judgment. *Science (80-. ).* **243**, 1668–1674 (1989).

19. Salganik, M. J. *et al.* Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* **117**, 8398–8403 (2020).

20. Kaplan, A., Skogstad, A. L. & Girshick, M. A. The Prediction of Social and Technological Events. *Public Opin. Q.* **14**, 93 (1950).

21. Mellers, B. *et al.* Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspect. Psychol. Sci.* **10**, 267–281 (2015).

22. Erev, I. *et al.* A choice prediction competition: Choices from experience and from description. *J. Behav. Decis. Mak.* **23**, 15–47 (2010).

23. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science (80-. ).* **211**, 1390–1396 (1981).

24. Tetlock, P. E., Mellers, B. A. & Scoblic, J. P. Bringing probability judgments into policy debates via forecasting tournaments. *Science (80-. ).* **355**, 481–483 (2017).

25. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**, 679–688 (2006).

26. Jebb, A. T., Tay, L., Wang, W. & Huang, Q. Time series analysis for psychological research: examining and forecasting change. *Front. Psychol.* **6**, (2015).

27. Luhmann, M. Using Big Data to study subjective well-being. *Curr. Opin. Behav. Sci.* **18**, 28–33 (2017).

28. Schwartz, H. A. *et al.* Predicting individual well-being through the language of social media. in *Biocomputing 2016* 516–527 (World Scientific, 2016). doi:10.1142/9789814749411\_0047

29. Kiritchenko, S., Zhu, X. & Mohammad, S. M. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **50**, 723–762 (2014).

30. Witters, D. & Harter, J. *In U.S., Life Ratings Plummet to 12-Year Low*. (2020).

31. Axt, J. R. The Best Way to Measure Explicit Racial Attitudes Is to Ask About Them. *Soc. Psychol. Personal. Sci.* **9**, 896–906 (2018).

32. Nosek, B. A. *et al.* Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur. Rev. Soc. Psychol.* **18**, 36–88 (2007).

33. Hehman, E., Flake, J. K. & Calanchini, J. Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents. *Soc. Psychol. Personal. Sci.* **9**, 393–401 (2018).

34. Ofosu, E. K., Chambers, M. K., Chen, J. M. & Hehman, E. Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proc. Natl. Acad. Sci.* **116**, 8846–8851 (2019).

35. Charlesworth, T. E. S. & Banaji, M. R. Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychol. Sci.* **30**, 174–192 (2019).

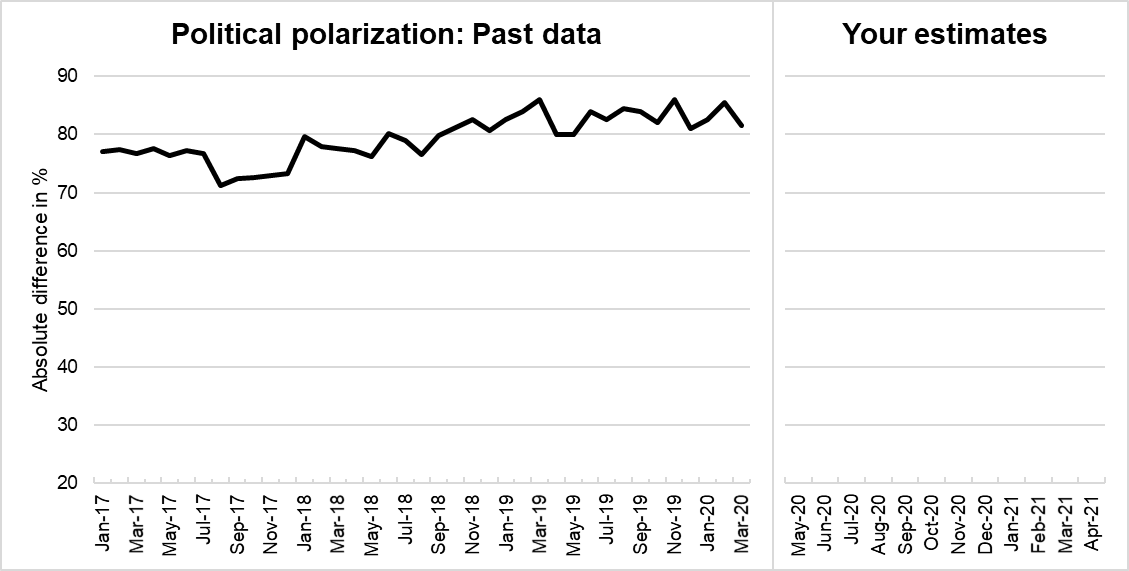
36. Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* **85**, 197–216 (2003).

37. Richard, F. D., Bond Jr., C. F. & Stokes-Zoota, J. J. One Hundred Years of Social Psychology Quantitatively Described. *Rev. Gen. Psychol.* **7**, 331–363 (2003).

**Appendix**

Example submission forms:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | |  | | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
|  | |  | |  |  |  |  | |  |  |  |  |
| Time | | Democrat | | Republican |  |  |  | |  |  |  |  |
| May-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Jun-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Jul-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Aug-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Sep-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Oct-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Nov-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Dec-20 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Jan-21 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Feb-21 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Mar-21 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| Apr-21 | |  | |  | **←** | **enter your monthly estimates in the box to the left** | | | | | |  |
| *Data Source:* | | Aggregated weighted data from the Congressional Generic Ballot polls conducted between January 2017 and March 2020. | | | | | | | | | | |
|  | | Congressional generic Ballot asks representative samples of Americans to indicate which party they would support in an election. | | | | | | | | | | |
|  | | Obtained from projects.fivethirtyeight.com/congress-generic-ballot-polls | | | | |



|  |  |  |  |
| --- | --- | --- | --- |
| Time | Polarization Score |  |  |
| May-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Jun-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Jul-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Aug-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Sep-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Oct-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Nov-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Dec-20 |  | **←** | **enter your monthly estimates in the box to the left** |
| Jan-21 |  | **←** | **enter your monthly estimates in the box to the left** |
| Feb-21 |  | **←** | **enter your monthly estimates in the box to the left** |
| Mar-21 |  | **←** | **enter your monthly estimates in the box to the left** |
| Apr-21 |  | **←** | **enter your monthly estimates in the box to the left** |

|  |  |
| --- | --- |
| *Data Source:* | Aggregated data from the Gallup polls conducted between January 2017 and March 2020. |
|  | Data represents absolute value of the difference score in Presidential Job Approval by Party Identification (Democrat vs. Republican) |
|  | Obtained from news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx |