On the Accuracy, Media Representation, and Public Perception of Psychological Scientists' Judgments of Societal Change.

Cendri A. Hutcherson[1,2], Konstantyn Sharpinskyi[3], Michael E. W. Varnum[4], Amanda Rotella[3], Alexandra S. Wormley[4], Louis Tay[5], and Igor Grossmann[3]

[1] Department of Psychology, University of Toronto Scarborough, Toronto, ON M1C 1A4, Canada
[2] Department of Marketing, Rotman School of Management, University of Toronto, Toronto, ON M5S 3E6, Canada
[3] Department of Psychology, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
[4] Department of Psychology, Arizona State University, Tempe, AZ, 85287
[5] Department of Psychological Sciences, Purdue University, West Lafayette, IN, 47907

in press

at

American Psychologist

*Correspondence should be addressed to Igor Grossmann, PAS 3047, University of Waterloo, Waterloo, ON N2L 3G1, igrossma@uwaterloo.ca..

Author Contributions: C. H. and I.G. developed the study concept. C.H., K. S., M.E.W.V., A.R, A.W. and I.G. designed the research. A.R., K.S. and I.G. collected the data. K.S., C.H., and I.G. analyzed the data. A.W. cross-validated the code and data analysis. C.H., K.S., M. E.W.V. and I.G. drafted the manuscript. L. T. provided critical feedback on data analyses. All authors read and provided feedback on the manuscript.

The authors have no known conflict of interest to disclose.

Abstract

At the onset of the COVID-19 pandemic, psychological scientists frequently made on-the-record predictions in public media about how individuals and society would change. Such predictions were often made outside these scientists' areas of expertise, with justifications based on intuition, heuristics, and analogical reasoning (Study 1; N = 719 statements). How accurate are these kinds of judgments regarding societal change? In Study 2, we obtained predictions from scientists (N = 717) and lay Americans (N = 394) in the spring of 2020 regarding the direction of change for a range of social and psychological phenomena. We compared them to objective data obtained at six months and one year. To further probe how experience impacts such judgments, six months later (Study 3), we obtained retrospective judgments of societal change for the same domains ($N_{scientists}$ = 270; $N_{layPeople}$ = 411). Bayesian analysis suggested greater credibility of the null hypothesis that scientists' judgments were at chance on average for both prospective and retrospective judgments. Moreover, neither domain-general expertise (i.e., judgmental accuracy of scientists compared to laypeople) nor self-identified domain-specific expertise improved accuracy. In a follow-up study on meta-accuracy (Study 4), we show that the public nevertheless expects psychological scientists to make more accurate predictions about individual and societal change compared to most other scientific disciplines, politicians, and non-scientists, and they prefer to follow their recommendations. These findings raise questions about the role psychological scientists could and should play in helping the public and policymakers plan for future events.

Keywords: scientific intuitions; science communication; COVID-19; forecasting; lay theories of change

Public Significance Statement

At the onset of the COVID-19 crisis, psychological scientists contributed to the public discourse on COVID-related societal change in the news media through intuition-based reasoning, and often

made predictions outside their area of expertise. We assessed the likely accuracy of such judgments by surveying psychological scientists and laypeople at the onset of the pandemic regarding future societal change in different domains and comparing predictions to actual markers of change at six months and one year after. We found that psychological scientists and laypeople made similar and largely inaccurate predictions. Neither direct experience, training nor domain-specific expertise was associated with greater accuracy.

Author's Note: This work was supported by a Social Sciences and Humanities Research Council of Canada Insight Grant xxx-xxxx-xxxx (to [anon.]), a National Science Foundation grant xxxxxx (to [anon.]), and a Social Sciences and Humanities Research Council of Canada Insight Grant xxx-xxxx-xxxx and Early Researcher Award ERxx-xx-xxx from the Ontario Ministry of Research and Innovation (both to [anon.]). Data and study materials, along with code for reproducible analyses are available at https://osf.io/9btsy/?view_only=ae9ab59aeb344e408c364beebc744385. This study's design, data exclusions, and portions of the analysis were preregistered; see https://osf.io/zxavd/?view_only=58483933208d4b2aaaf49d82d7ca057a. The authors gratefully acknowledge Alden Lai and the Gallup-WPE Global Wellbeing Initiative for providing estimates for well-being, Giving Tuesday for providing estimates for charity domains, and Lynn Vavreck and the Nationscape initiative for providing data on political polarization.

Formal training in the social sciences typically focuses on developing explanatory theories that account for observed, often laboratory-based, phenomena. Although such approaches have resulted in richly detailed causal models of individual human behavior, recent years have witnessed growing calls to understand whether and how to increase their usefulness (Watts, 2017). Can psychological theory scale up to predict larger societal processes (Yarkoni & Westfall, 2017) in ways that enable effective intervention in times of crisis (IJzerman et al., 2020)? At the onset of the COVID-19 pandemic, there were a handful of notable efforts by psychologists and other behavioral and social scientists to provide formal guidance in academic journals about what areas of individual and societal behavior might be affected (Brooks et al., 2020; van Bavel et al., 2020; West et al., 2020). Scientists also attempted to contribute to public understanding of the potential consequences of the pandemic through discussion in public media, such as newspapers and magazines. How often and on what basis did psychological scientists make those judgments? For example, are such public judgments grounded in a more intuitive and heuristic reasoning style, or are they based on some combination of expert knowledge and formal modeling of potential outcomes? Are such judgments of societal change accurate? Here, we ask how psychological scientists made on-the-record judgments about societal change in public media, and formally assess whether the nature of their expertise gives them an advantage in the accuracy[1] of their judgments about future outcomes, compared to an average non-expert.

Understanding how psychological scientists make public judgments is critical for determining their accuracy and potential usefulness. On the one hand, psychology training and expertise should improve understanding of probability and statistics (Fong & Nisbett, 1991;

[1] Our main operationalization of accuracy concerns prediction of direction of societal change, because most psychological causal models of human behavior or social processes lend themselves to predictions about direction of change (e.g., "if *X* occurs, violence will increase") rather than estimates of specific magnitude (e.g., "if *X* changes by *Y* amount, violence will increase by *Z* amount). Supplementary Results show similar conclusions when accuracy is operationalized via magnitude or the rank ordering of change across different domains.

Nisbett et al., 1987) and reduce mistaken assumptions about human behavior (Gardner & Dalsing, 1986; Gardner & Hund, 1983; Taylor & Kowalski, 2004)—qualities that tend to increase the accuracy of forecasts for discrete geopolitical events (Mellers et al., 2015). Furthermore, the existence of empirically-grounded, causal theories about human responses to social isolation (Hawkley & Cacioppo, 2010), financial uncertainty (Artazcoz et al., 2004), and disease threat (Schaller & Park, 2011) should enable psychologists to estimate, at minimum, the direction of changes in psychology and behavior in response to the pandemic. On the other hand, research on forecasting in domains ranging from political (Tetlock, 2005) and economic (Armstrong, 1985), to career-related outcomes (Ægisdóttir et al., 2006; Dawes et al., 1989) suggests that experts are rarely more accurate than simple statistical models (Tetlock, 2005). Moreover, psychological theories are typically applied only at the individual or local level. Different forces may be at play when generalizing to societal processes writ large (Na et al., 2010; Piantadosi et al., 1988).

To better understand the nature of psychological scientists' contribution to public understanding in times of crisis, we first analyzed the world's largest corpus of COVID-19 news reports, tracing the nature of psychological scientists' engagement with the new media during March-May 2020, finding that these judgments were typically made in an intuitive style, relying on analogical reasoning and only occasionally on reference to research, and that more than a quarter of judgments were done outside of scientists' domain-specific expertise (Study 1). Then, we present a systematic investigation into the accuracy of such judgments (in both absolute terms and relative to laypeople) for predicting and retrospectively evaluating aggregate-level changes in human psychology and behavior during the first six months of the pandemic (Studies 2-3). We find that for most domains, scientific judgments of the kind found in public discourse were either at chance or largely inaccurate, and not more accurate than judgments of laypeople. Domain-specific expertise did not facilitate prediction accuracy. Finally, we show that understanding psychological scientists' accuracy matters because the public expects them to be more accurate

in predicting psychological and societal outcomes, and prefers to base policy on their recommendations, compared to politicians, laypeople, and other scientific disciplines (Study 4).

Methods

The project was approved by the Office of Research Ethics at the University of Waterloo (#42123 & #43189). Pre-registration, materials, methods, code, and reproducible analyses are available on the Open Science Framework at https://osf.io/9btsy/?view_only=ae9ab59aeb344e408c364beebc744385.

Study 1: News media engagement of psychological scientists

In Study 1, we examined how psychological scientists talked about the pandemic in the news media. To this end, we used The Coronavirus Corpus (https://www.english-corpora.org/corona/) – an extensive record of over 1.8 million texts appearing in online newspapers and magazines in 20 different English-speaking countries – to identify online articles in newspapers and magazines that contained an interview with an academic psychologist regarding some aspect of the pandemic. To identify candidate articles, we first located texts containing key search terms (e.g., psychologist, psychology professor, psychology researcher), limited to a publication date between March 15, 2020 and May 15, 2020. From this, we identified a subset of texts containing interviews with psychological scientists about the pandemic (see Figure S1 for a flowchart detailing the definition and identification of psychological scientists). These articles were then reviewed by hand to remove duplicates, and to apply additional exclusions (e.g., not actually containing an interview with a psychologist despite containing key search terms). This produced a database of 169 unique English-language articles appearing in a wide variety of outlets, including the *New York Times*, *Wall Street Journal*, and other high-quality news sources, presenting judgments by 213 individual scientists. Because these scientists frequently commented on multiple distinct topics (e.g., effects of the pandemic on depression and also children's cognitive development), this yielded 719 unique judgments about the

consequences of the pandemic, which were coded by three independent raters for whether the expert's judgment on a given topic fell within their particular domain of expertise, whether it was an observation about the present or a prediction for future outcomes, and what type of justification was given (i.e., none, current events, historical analogy, research, other), as well as the certainty of language used (interrater agreement 79-85%, see Supplementary Methods for details).

Study 2a: Psychological scientists' predictions about societal change

Participants

In the first two days of April 2020, we recruited psychology experts by circulating a call for forecasts on listservs and mailing lists for the Society of Personality and Social Psychology (SPSP), the Cognitive Science Society (CogSci), Society for Research in Child Development Commons, Association for Behavioral and Cognitive Therapies, and the Society for Judgement and Decision Making (JDM). We also posted in relevant Facebook groups, including Psychological Methods, PsychMAP, and COVID-19 groups. Additionally, we contacted colleagues and graduate students at the authors' affiliated departments and institutes.

A total of 470 scientists provided their forecasts in April. Of these, six had incomplete responses, four participants provided nonsensical responses (e.g., age > 900), and 57 participants answered all survey questions in less than five minutes (pilot testing with research associates revealed that five minutes is the minimum necessary time to complete the study), and two participants indicated they were undergraduate students. These responses were removed. The final sample (N = 401) consisted of participants from 39 countries, with demographics that closely match the membership of psychological societies relevant to these predictions (see Table S1 in the Supplemental Online Materials).

Procedure

Participants first answered several demographic questions. Participants next predicted cultural change in the U.S. for 11 domains, presented in a randomized order: implicit and explicit prejudice towards minorities, political polarization, traditionalism, individualism, generalized trust,

delay of gratification, expected birth rates, concern for climate change, life satisfaction, and clinical depression (see verbatim questions on Open Science Framework at osf.io/npzcr). Participants provided predictions for six months, one year, and two years in the future on a sliding scale ranging from 50% or greater decrease (-50) to 50% or greater increase (+50). Of these 11 domains, we were able to obtain reliable benchmarks to assess accuracy for seven: polarization, traditionalism, individualism, trust, climate change, life satisfaction, and depression (see Accuracy Analyses section below).

Beyond the eleven domains we provided to participants to make forecasts, we were interested in participants' unstructured views about the key societal domains in which one might observe significant changes. After participants predicted cultural change in the above variables, we asked them to identify one key psychological or social issue in the United States not covered in the survey that they thought would change.

Study 2b-c

Participants

Whereas Study 2a focused on predictions before the initial peak of COVID-19 cases in the U.S., Studies 2b and c were conducted after the initial peak. In the last week of April and the first week of May 2020, we recruited another group of psychological scientists using the same methods described in Study 2a. A total of 354 psychological scientists provided their forecasts during this time (98% non-overlapping with the early April sample). Of these, we removed two who had incomplete responses, 31 who completed the survey too fast according to pilot test estimates (< 5 minutes), and four who indicated they were undergraduate students. The final sample included 316 participants from 26 countries (see Table S1 for demographic information).

Concurrently in the first week of May 2020, we also obtained forecasts from a nationally representative sample of English-speaking U.S. residents via the crowdsourcing UK-based company Prolific (www.prolific.co, Study 2c). To recruit a nationally representative sample, Prolific

uses the intended sample size (target *N* = 400) to stratify across age, sex, and ethnicity, based on census data from the U.S. Census Bureau (Prolific, 2020). Of the 411 participants who attempted the study, we removed three who had incomplete responses and 14 who completed the survey in less than 5 minutes. The final sample consisted of 394 participants. Prolific participants received 1.25USD for completing the survey.

Procedure

Participants in Study 2b and 2c followed the same general procedure outlined for Study 2a, with the following differences. In addition to the 11 domains of Study 2a, they made predictions in 4 additional domains: loneliness, religiosity, charitable giving, and prevalence of violent crimes (verbatim questions on osf.io/npzcr and in Table S1 in the Supplemental Online Material). For each domain, participants made predictions as in Study 2a, but also rated confidence in their predictions on a 5-point scale (1 = Not at all to 5 = Extremely). Participants also answered additional demographic questions (see Supplemental Online Material for verbatim items). Of these 15 domains, we were able to obtain reliable benchmarks to assess accuracy for 10: loneliness, charitable giving, violent crimes, polarization, traditionalism, individualism, trust, climate change, life satisfaction, and depression. Thus, in the remainder of the paper, we focus on responses from our participants in these domains (see the Accuracy Analyses section below). Although this number of domains is not extensive, we think it likely represents a best-case scenario for assessing the utility and accuracy of psychological scientists' forecasts, since these are the domains for which a) psychology has established theories about how change might occur in the face of the pandemic; b) there was sufficient interest that high-quality data was being measured during the pandemic; and c) psychologists were generally more likely to comment in the media. However, we acknowledge that all conclusions we make come with caveats due to the limited number of domains, and apply largely to domains in which psychology makes straightforward pandemic-related predictions rather than all possible judgments in general.

Study 3a-b: Psychological scientists' retrospective judgments about societal change

In Study 3a and 3b, we aimed to compare prospective predictions from Study 2 to retrospective estimates of changes in these same domains. Study design and data exclusions were preregistered (registration available at https://osf.io/zxavd/?view_only=58483933208d4b2aaaf49d82d7ca057a).

Participants

We recruited a new regionally-stratified sample of Americans from Prolific. Participants received 1.10 GBP for participation. Exclusion criteria were identical to Study 2, with the exception that we also preregistered to exclude participants who provided estimates for fewer than five domains, or indicated at the end of the survey they took part in the April/May prediction studies, even though there was no April survey for Prolific and none of the Prolific IDs from May, 2020 survey match their Prolific IDs). Of the 445 participants who started the study, we removed 27 who had incomplete responses and seven who indicated they took a forecasting survey in April. The final sample consisted of 411 participants.

We also obtained survey responses from a sample of psychological scientists, recruited via mailing lists (e.g., Social and Personality Psychology mailing list, JDM mailing list) and social media. Similar exclusion criteria were applied to this sample, with the exception that we did not require scientists to be U.S. citizens. A total of 350 psychological scientists provided forecasts during the last week of October/first week of November 2020 (88% non-overlapping with the forecasting samples in Studies 1-2). Of these, we removed 80 responses because they provided fewer than five domain estimates. The final sample included 270 participants (see Table S1).

Procedure

Participants in Study 3 were asked to provide retrospective assessments of percentage change as well as confidence in their assessments for the same 15 domains as in Study 2b (verbatim questions on https://osf.io/9btsy/?view_only=ae9ab59aeb344e408c364beebc744385

and additional information in Supplemental Online Material). To match instructions in Study 2, participants were instructed to "provide an estimate of how much you think it has changed compared to where the issue stood six months ago (i.e., end of April 2020)." In addition, as an exploratory analysis, we obtained information about the types of information participants considered when making their judgments, including whether they considered specific news reports, or brought to mind vivid personal memories (see Supplemental Online Materials for detail). All other details were as in Study 2b.

Study 4: Lay perceptions of scientists

Participants

For Study 4a, we recruited a sample of Americans from Prolific in March 2021. Participants received 1.10GBP for participation. Of the 220 participants who started the study, we removed 11 who did not provide any responses and six who did not provide a comprehensible answer to an open-ended question at the end of the study. The final sample consisted of 203 participants (M_{age} = 33.81, SD_{age} = 13.04; 57% female/ 41% male/2% non-binary; 74% White/7% Latinx/8% Asian-American/6% Black/5% Other). To supplement these results, we also recruited a sample of academics and policymakers via announcements on social media (Study 4b). Thirty individuals filled out the survey (M_{age} = 40.32, SD_{age} = 10.64; 57% female/39% male/4% non-binary; 78% White/9% East Indian/9% Mixed/4% Other).

Procedure

Participants considered different groups of scientists, practitioners, and the layperson, and rated their possible accuracy when predicting societal change over COVID-19 for depression, life satisfaction, loneliness, violence and related domains, and who would they like to make recommendations for these societal issues. Participants were presented with three sets of questions concerning predictions, preference for recommendations, and ranking of the top three groups they would prefer to ask how the COVID-19 pandemic will affect human behavior and

society in the long term. For each set of questions, participants were presented with ten groups: scientists with expertise in psychology, economics, epidemiology, history, political science, or public health, practitioners with expertise in social work or medicine, as well as politicians and the average American. See the project's Open Science Framework page (https://osf.io/dr9a8) for the precise wording of questions.

We also examined whether participants read prior reports about behavioral science expertise for predicting societal trajectories over COVID-19. Only 7% of the sample indicated vague familiarity with such reports, and excluding these participants did not change the results.

Accuracy Analyses (Studies 2a-c and Studies 3a-b)

We targeted estimates for all domains where we could locate large-scale, nationally representative surveys assessing the state of that domain in April/early May and in October/early November. Our chief question concerned societal-level change. Thus, we relied on cross-sectional data as long as the estimates were sufficiently large and representative of the U.S. population at large. When possible, we used weighted averages to adjust for representativeness as per the U.S. Census. If we could locate multiple sufficiently representative indicators for a given domain, we performed parallel analyses with each. Our sources included the Household Pulse Survey from the National Center for Health Statistics and the U.S. Census Bureau, USC's Understanding America Survey, Nationscape, Gallup Panels, the National Commission on COVID-19, Criminal Justice and Giving Tuesday, among others. See Supplementary Table S2 for the exact wording of the questions, and Table S3 for more information on each source. When estimates were based on the percentage of the population at the given time point, we calculated the difference score. When the data was based on the sample estimate of a scale-based response, we calculated the percentage change between the initial estimate of the sample in April 2020 and the subsequent estimate half a year later. Ultimately, we quantified societal change in the U.S. for ten domains, with most estimates coming from nationally representative surveys and

aggregated official reports of crime. We report estimates for two additional benchmarks with lower sampling consistency (prejudice markers from Project Implicit) in the online supplement. In addition, we obtained objective benchmark data for four of these ten domains one year after the start of the pandemic (five surveys were no longer collecting data, preventing comparable accuracy analyses). Additionally, for one-year markers, we also obtained U.S. birth rate statistics from the Human Fertility Database (https://www.humanfertility.org/cgi-bin/main.php), a joint project of the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany and the Vienna Institute of Demography (VID) in Vienna, Austria.

Our main criterion for accuracy was the direction of change (increase/decrease) as a function of the type of estimate (prospective / retrospective), sample type (lay / expert) and domain type. In addition to frequentist statistics, we ascertained the strength of evidence for or against specific hypotheses about the accuracy of psychological scientists using an estimation of Bayes Factors (Rouder et al., 2009) provided by the function *bayesfactor_models* from the bayestestR package. In secondary analyses, we compared the magnitude of predicted change to observed change, including both average estimated change at six months, as well as the estimated trajectory of change over the full two-year prediction period. In addition, we performed a number of supplemental analyses quantifying accuracy at six months by the percentage of the sample falling within a certain range of observed change, as well as rank-order accuracy across domains (i.e., predicting which domains would show the *most* vs. *least* change). See additional results in the Supplemental Material for details.

Results

Study 1: Psychological scientists' judgments in the news media

To understand how psychological scientists' judgments might shape public perceptions, we first sought to understand *how* they typically make such judgments, and to document how the

topics they discussed aligned with their expertise. This analysis allowed us to answer a crucial question: when communicating to the public, how often do psychological scientists base their judgments on discipline-specific expertise, theory, and models, or instead use an intuitive or heuristic reasoning style that might be shared with non-experts?

To determine what experts were saying about the societal impact of COVID-19, whether they were making predictions, and how they made them, we analyzed the comments made by psychological scientists to the news media in the first two months of the pandemic (see Methods for details). Analysis of the frequency of content-related words (excluding generic terms like 'well,' 'if,' etc., and terms related to 'psychological scientist', which were used to identify the articles) indicated that experts spoke on a number of topics focused on mental health, well-being, and various social effects of the pandemic (Figures 1a and 1b). Though most interviews with experts concerned observations about the current effects of the pandemic (72% of cases), explicit forecasts about the pandemic's future consequences were also common (28% of cases). When talking to the news media, more than a quarter of judgments were made outside of scientists' area of expertise (27% of cases). We observed no evidence of a difference between scientists speaking within or outside their domain of expertise in the likelihood of making a prediction vs. an observation, $\chi^2(1, N = 717) = 1.02$, $p = .31$.

Finally, we assessed what justification/rationale psychological scientists provided for their judgments. We found that for a sizable fraction of statements (47%), no justification for the judgments was included. When a justification was provided, it rarely referenced research or scientific theory (21% of cases). In most cases (73%), scientists were quoted making intuitive reference to present events (e.g., noting the hoarding of toilet paper when justifying the influence of the pandemic on panic responses).

To determine whether this lack of scientific justification could be attributed simply to omission by journalists, we analyzed separately op-eds in which a psychologist spoke for

themselves rather than articles in which they were quoted by a journalist. Although op-eds were more likely to give any kind of justification for a judgment (68% of op-ed judgments versus 51% of quoted judgments), $\chi^2(1, N = 717) = 6.84$, $p = .01$, we found no evidence of a significant difference in the likelihood of that justification being based on research (19% of justifications in op-eds, 22% of justifications in other news articles), $\chi^2(1, N = 376) = 0.07$, $p = .80$. Thus, even when psychological scientists were given full control of the narrative via an op-ed format, justifications were either absent or merely reflected references to present-day events.

Across all article types, a significant difference emerged in the type of justification between scientists speaking within or outside their domain of expertise, $\chi^2(3, N = 717) = 9.63$, $p = .02$. Domain-experts' judgments were significantly more likely to reference research compared to scientists without domain expertise (13% of expert judgments versus 5% of non-expert judgments), $z = 3.75$, $p < .001$, whereas non-experts were equally likely to omit vs. provide justifications for their judgments (51% vs. 46%), $z = 1.05$, $p = .15$. However, when giving a justification, both domain experts and non-domain experts were more frequently quoted referencing current events than research (70% of justifications for domain experts, 82% of justifications for non-domain experts), both binomial tests $p < .001$.

Study 2a: The accuracy of psychological scientists' spontaneous judgments

Our analysis of the types of judgments made by psychological scientists in the news media suggested that these judgments might often be made on the spot, without an extensive rationale, or with an intuitive rather than empirical basis for judgment. This observation raises a question about the accuracy of judgments that psychological scientists conveyed to the media in the aftermath of the COVID-19 pandemic.

To address this question, we analyzed predictions about outcomes of the pandemic in the United States from two samples of psychological scientists, one collected in early April 2020 (Study 2a; $N = 401$) and another collected in late April/early May 2020 (Study 2b, $N = 316$).

Scientists could make these predictions however they chose, including formal model analysis. Survey completion times and post-hoc analysis of self-reported strategies suggest that the majority likely relied on spontaneous, intuitive judgments informed by both training and life experiences, similar to what we observed in news media quotes (see Online Supplemental Results for details), although we acknowledge that such interpretation is speculative. Predictions were obtained about change in different domains (e.g., depression, political polarization) at six months, one year and two years into the future (see Figures S2 and S3 for predicted trajectories). Although we aimed primarily to recruit psychological scientists (composing ~80% of the sample), we also attracted responses from individuals in other behavioral science disciplines, such as economics, political science, and sociology, allowing us to compare psychological and non-psychological disciplines. However, these analyses indicated little consistent distinction among disciplines on predictions or accuracy (see Table S12-13, Figure S12 and Online Supplemental Results). We thus report statistics for the full sample here, focusing on other definitions of expertise (e.g., domain-specific training, level of education) as potential moderators of accuracy.

In each of these surveys, we asked our participants to consider specific domains for which a sizable body of theoretical and empirical work links these variables to pathogen-related threats. We focus here on domains for which we could obtain high-quality, national-level data. Based on theories that suggest that intergroup processes are affected by evolutionary and ontogenetic pressures related to pathogen stress (Faulkner et al., 2004; Fincher et al., 2008; Murray et al., 2011; Schaller & Park, 2011; Tybur et al., 2016), we examined judgments of political polarization, cultural values related to traditionalism and individualism, as well as prosocial and antisocial behavior. Based on life history theory, which argues that organisms increase present-focused behavior and reproduction in response to environmental threat and pathogen-related unpredictability (Griskevicius et al., 2011; Horn, 1978), we assessed birth rates. Finally, based on theories about how human mental and affective well-being is influenced by stressors (Kendler et

al., 1999), including social isolation (Hawkley & Cacioppo, 2010), we assessed judgments of depression, loneliness, and life satisfaction.

To assess the accuracy of such judgments, we compared predictions for six months to ground truth markers of change at six months for depression, life satisfaction, generalized trust, loneliness, individualism, traditionalism, political polarization, climate change attitudes, violent crimes, charitable giving. At 12 months, we were able to obtain high-quality ground truth markers for five domains: depression, loneliness, birth rates, violent crimes, and charitable giving (see also estimates for explicit and implicit prejudice in the online supplement for both 6 and 12 months).

Were scientists more accurate than chance in predicting societal change across domains?

We answer this question by comparing scientists' predictions for 6 and 12 months into the pandemic against ground truth. At the six-month mark, we examined the intercept term of a mixed-effects logistic regression (see Supplemental Online Materials for detail) with directional accuracy (1 = correct / 0 = incorrect) in each domain as the dependent measure and participant (total N = 707) as a random intercept. This yielded an average individual accuracy of 50.5% [49.0 51.7], a value that was not significantly different from chance, $z = 0.61$, $p = .54$.

We then investigated how accuracy varied by domain at six months. As Figure 2 (top panel) shows, scientists showed above-chance directional accuracy in only four out of ten domains at six months. They correctly predicted increases in depression (89% correct), binomial test against change accuracy of 50% $p < .001$, corrected for multiple comparisons; political polarization (73% correct), $p < .001$; and charitable giving (59% correct), $p = .03$; and correctly predicted decreases in generalized trust (61% correct), $p < .001$. However, in most of these domains, scientists tended to significantly overestimate the magnitude of changes, all t-tests against actual change > 2.77, .006 < ps < .001, corrected for multiple comparisons. The only exception concerned charitable giving, t-test against actual change = 1.55, $p = .12$ uncorrected. Moreover, for the remaining six domains, they either failed to predict direction above chance levels or were actually significantly *worse* than chance. They incorrectly predicted decreases in

life satisfaction (13% correct), binomial test $p < .001$; loneliness (17% correct), $p < .001$; individualism (35% correct), $p < .001$; and concern for climate change (42% correct), $p < .001$, and were no better than chance in predicting changes in traditionalism (50% correct), $p = .97$, and violence (54% correct), $p = .27$. Conclusions did not change when using alternative measures of accuracy, such as absolute deviation or rank ordering of the magnitude of changes across domains (see Online Supplemental Results for details). Conclusions were also similar when making a more granular comparison of early and late April predictions with early and late October markers, respectively (see Online Supplemental Results for relevant details).

Scientists' average individual accuracy for 12-month predictions (Figure S9) was even worse than at six months and substantially lower than chance, 35.1% [33.3, 37.1], $z = -13.58$, $p < .001$. Moreover, when investigating accuracy by domain, psychological scientists made accurate *directional* predictions for only two out of the five domains for which we had objective markers: birth rate (54% correct) and violence (65% correct), ps < .02. However, they nonetheless either over- or underestimated the *magnitude* of change in these domains, ps < .001. In addition, directional accuracy for the remaining three domains was significantly worse than chance, all ps < .001 (depression - 9% correct, loneliness - 21% correct, and charity - 38% correct).

Study 2b: Expert predictions are not more accurate than lay predictions

Although psychological scientists' judgments were largely inaccurate overall, it could still be the case that these predictions were more accurate on average than those of laypeople. To test this possibility, we collected predictions from a nationally stratified sample of Americans (N = 394) in late April/early May of 2020 in parallel with the collection of our second sample of psychological scientists. We then compared the directional accuracy of lay predictions to psychological scientists. Results of a generalized linear mixed model with accuracy of prediction scores for direction of change as a dependent variable and expertise as a predictor revealed no evidence for a difference between psychological scientists and laypeople, χ^2 (1, N = 1,101) = 2.14, $p = .14$. Moreover, comparison of a model including group (scientist vs. lay) as a factor to a

model without this factor yielded a Bayes Factor (Rouder et al., 2009) of 95 in favor of the null. Thus, the evidence increased credibility of the null hypothesis that there was no advantage for scientists over lay people in predictive accuracy, at least in the kinds of domains examined here.

Study 3a-b: Retrospective judgments were not more accurate than prospective ones

Our results suggest that the prediction of large-scale trends in psychological and societal outcomes might be difficult for both scientists and laypeople alike, and that scientists were similar in accuracy to the average American. However, this could occur for many reasons related to chaotic or unpredictable dynamics in response to the pandemic. We reasoned that if experts mispredicted the effects of the pandemic solely due to unforeseeable dynamics, but would otherwise have made more accurate judgments about how the pandemic affects psychological outcomes, then expert judgments of change should be more accurate in retrospect, especially compared to laypeople. In other words, experts should be better able to update their judgments in light of experience and/or direct observation of empirical data (although, for most domains, this data did not yet exist or was not yet published, likely leaving most scientists to rely on the same kinds of intuitive experiences and knowledge as laypeople).

To assess whether this was the case, we conducted a set of pre-registered surveys (https://osf.io/9btsy/?view_only=ae9ab59aeb344e408c364beebc744385) in a third sample of psychological scientists (Study 3a, N = 270) and a nationally-stratified sample of lay Americans (Study 3b, N = 411) in late October/early November, just before to the U.S. election. We asked participants to estimate how much change *had already* occurred in the previous six months, rather than to make forecasts of future change.

Our results suggested that retrospective judgments improved slightly compared to scientific predictions half a year prior (Figure 2, bottom). On average across all domains, psychological scientists had an accuracy rate in retrospective reports of 51.9% [50.0, 53.7], , which was slightly but significantly more accurate than prospective reports, odds ratio = 1.25, z = 3.35, p = .001. However, the domains in which psychological scientists' prospective judgments

were inaccurate were the same domains in which their retrospective judgments were inaccurate. Moreover, in domains where most predictions were inaccurate, even larger numbers of retrospective assessments were directionally inaccurate, $\chi^2(1, N = 1{,}782) = 150.51$, $p < .001$, in large part because predictions became more extreme (see Online Supplemental Results). Finally, despite the fact that psychological scientists' accuracy improved in retrospective reports compared to prospective ones, this improvement was non-significantly *smaller* than that of laypeople, $\chi^2(1, N = 1{,}782) = 0.84$, $p = .39$, Bayes Factor = 83 in favor of the null.

Domain-specific expertise was not linked to greater accuracy

Although we did not find any difference between the average scientist and the average layperson, it is possible that experts with extensive training in a specific topic might perform better. Since these are the scientists most likely to be consulted both by the media and public policymakers, it is important to know whether they provide more accurate estimates within their specific knowledge area. We thus examined whether domain-specific expertise was associated with greater accuracy. To do this, we conducted regression analyses asking whether directional accuracy was significantly different when made within or outside a domain in which an expert self-reported having expertise or training (see Supplemental Methods for coding of domain-specific expertise in each study). Operationalizing expertise this way, we found no significant effect of expertise for either prospective predictions, $\chi^2(1, N = 659) = 0.06$, $p = .81$, or retrospective estimates, $\chi^2(1, N = 270) = 0.20$, $p = .66$. We also asked whether the degree of experience more generally (i.e., graduate student, post-doc, untenured, tenured faculty) mattered. Although we did observe an effect of experience on prediction accuracy in specific domains, $\chi^2(18, N = 581) = 31.54$, $p = .03$, this was largely driven by an advantage for *graduate students* over faculty in a small set of domains (see Online Supplemental Results for details). In other words, greater expertise did not seem to confer special ability to consistently and correctly predict outcomes.

Scientists are less confident in their estimates than laypeople

We did observe one notable difference between psychological scientists and lay Americans: scientists were consistently less confident in both their predictions, $5.14 < zs \leq 9.86$, all $ps < .001$, and their retrospective estimates, $4.77 < zs \leq 9.04$, $ps \leq .001$[2]. Greater confidence was simultaneously associated with a greater probability of correctly estimating the direction of societal change, $\chi^2(1, N = 1{,}387) = 6.02$, $p = .01$, but over-estimating its *magnitude*, $\chi^2(1, N = 1{,}387) = 88.62$, $p < .001$. As Figure 3 shows, these effects were each magnified in retrospective compared to prospective estimates, such that confidence corresponded to fewer errors in directional inaccuracy, $\chi^2(1, N = 1{,}387) = 5.64$, $p = .02$, but larger errors of magnitude, $\chi^2(1, N = 1{,}387) = 110.29$, $p < .001$, in retrospective compared to prospective estimates. Thus, for both scientists and lay individuals, predictions made with greater confidence were more likely to get the direction of change correct, yet also to overestimate its magnitude.

Sources of information when making judgments about societal change

To understand the kinds of information scientists and laypeople used in constructing their judgments, in Study 3 we asked participants whether they relied on vivid personal experiences and/or news reports when estimating societal change in the last six months (see Supplemental Methods for details). Both lay individuals and psychological scientists were more likely to report relying on news reports (scientists = 45% of judgments, lay individuals = 41% of judgments) than personal experiences (scientists = 30% of judgments, lay individuals = 30% of judgments), scientists: $\chi^2(1, N = 270) = 121.32$, $p < .001$, lay individuals: $\chi^2(1, N = 411) = 78.03$, $p < .001$. This difference was somewhat larger among scientists, $\chi^2(1, N = 681) = 5.53$, $p = .02$, due to scientists reporting non-significant trends to rely more on news reports, $\chi^2(1, N = 681) = 2.91$, $p = .09$, and less on personal experience, $\chi^2(1, N = 681) = 0.92$, $p = .34$. Thus, consistent with the

[2] The effect also held when controlling for political affiliation, ethnicity, age, gender and income, $3.05 < Zs \leq 7.24$, $.002 < ps \leq .001$.

observation that psychological scientists and laypeople did not differ in their estimates or in the accuracy of their estimates, scientists and laypeople showed largely similar use of non-scientific sources of information. Intriguingly, relying on concrete personal experiences was associated with greater directional accuracy, $\chi^2(1, N = 681) = 26.75$, $p < .001$, while relying on news articles was associated with lower accuracy, $\chi^2(1, N = 681) = 4.60$, $p = .03$. Nevertheless, effects on accuracy of considering personal experience and news were similar for scientists and lay individuals: personal experience, $\chi^2(1, N = 681) = 3.09$, $p = .08$; news, $\chi^2(1, N = 681) = 0.83$, $p = .36$.

Study 4: The public's preference for expert judgments

One might argue that the accuracy of psychological scientists' estimates of societal change matters chiefly if the public actually values such pronouncements, or would prefer to hear from psychologists as opposed to other sources (e.g., medical practitioners or politicians). We examined the latter hypothesis by surveying a sample of U.S. residents ($N = 203$) about who they expected to be most accurate in predicting societal trends in depression, well-being, violence, and related domains over the first half year of the crisis, and who they would most prefer to consult about how the COVID-19 pandemic would affect human behavior and society (see Methods for details). As Figure 4 indicates, participants consistently ranked psychological scientists at the top of a list of different experts and practitioners, with economists and political scientists in the middle, and politicians ranked below even the average American. Psychological scientists were viewed as significantly more accurate than most groups, $2.37 < t\text{s} \leq 20.93$, all $p\text{s} < .02$, false-discovery-rate (FDR)-corrected, and significantly more-preferred to provide recommendations than most groups, $3.50 < t\text{s} \leq 22.47$, $p\text{s} < .001$, FDR-corrected, with the exception of public health, which evoked similar levels of preference (see Supplemental Online Materials for further detail).

Discussion

Many psychological scientists were willing to comment publicly on the likely outcomes of the pandemic, justifying their analyses largely based on intuitive reasoning rather than a reference

to theoretical or quantitative models (Study 1). Yet such snap judgments about societal change in the wake of the COVID-19 pandemic were similar to those of laypeople (Studies 2a, 2b). The small improvements in accuracy that we observed in retrospective judgments were no larger than for laypeople (Study 3a, 3b). Nor did we find that scientists with greater scientific training, higher career stage, or domain-specific expertise—i.e., the individuals most likely to be consulted by both news outlets or policymakers—were more accurate. We also observed some evidence that the inaccuracy of judgments reported in the news might matter: among both scientists and lay individuals, those who reported relying on news reports when making judgments of change that had occurred over the first six months of the pandemic were significantly *less* accurate. These findings stand in contrast to the observation that psychological scientists are *believed* to be more accurate in predicting the pandemic's societal impacts compared to scientists in other disciplines, policymakers or the lay public (Study 4). Although our conclusions are limited by the small number of specific domains assessed here, they nevertheless suggest that scientists may use their greater knowledge of research and theory to justify, rather than shape, the intuitions they share with the average person.[3] This work raises important questions about how to improve the accuracy of scientists' predictions regarding the societal effects of events like the COVID-19 pandemic.

Improving scientific accuracy requires some understanding of why psychological scientists were no more accurate than laypeople at predicting the pandemic's societal consequences. We propose two interrelated explanations. First, most psychological scientists have little training in

[3] It is possible that psychologists may have been more accurate if a larger or different set of domains had been chosen. In the present work we were limited to a relatively small set of domains by four factors: 1) we only selected domains where there was prior reason to anticipate substantial change as a result of the pandemic, 2) we only selected domains in which psychological scientists were likely to have some knowledge or expertise, 3) we only selected domains with high quality national datasets, and 4) we could only obtain a limited set of predictions from the psychological scientists participating in our study, who were volunteers with limited time (a pragmatic concern that was especially relevant during the initial pandemic lockdowns). We do not see a clear reason why such judgments would have been more accurate for a different or larger suite of domains, but we remain open to the possibility and hope that others might consider such an ambitious undertaking with more domains in the future.

prediction-oriented (as opposed to explanation-oriented) designs and models (Hofman et al., 2017; Yarkoni & Westfall, 2017). The fact that not only the predicted direction, but also the magnitude of societal change judgments aligned closely with those of the general public supports this interpretation. That there were no major differences in accuracy between graduate students and tenured faculty further corroborates the absence of benefits for experience in psychological science on such judgments. This lack of attention to out-of-sample prediction may limit generalizability of existing psychological theories that experts may draw on to estimate effects in the real world (Hofman et al., 2017; Yarkoni & Westfall, 2017). Second, although psychologists might be experts at making conditional statements about how and why pandemic-related behaviors might change *if* specific manipulations or policies were adopted (Ruggeri et al., 2022), formal psychological models of overall societal change in response to a once-in-a-century event like the pandemic are lacking (Ackerman et al., 2021). Without theory and necessary training to guide them, psychological scientists likely based their estimates on the same naïve theories of human social dynamics as laypeople (Heider, 1958; Kelly, 1955). Indeed their judgments were strikingly similar. Thus, we suspect that when scientists make intuitive judgments of the sort that we assessed here, and that appear in news media, they likely rely on exactly the same heuristics and reasoning as lay individuals.

One might also argue that perhaps expert predictions were inaccurate because policymakers heeded their cautionary advice and took actions that mitigated the negative outcomes that psychological scientists predicted. However, if this had been the case, then we likely should have observed greater differences between prospective and retrospective judgments (and reduced extremity of retrospective judgments), especially in domains like depression and subjective well-being, where policy responses might have had the greatest impact. Instead, we find that retrospective judgments in these particular domains are generally *more* inaccurate and extreme.

Our findings also suggest that the level of analysis at which psychologists generally excel (i.e., predicting behavior of individuals or small groups) may not prepare them to provide judgments at a higher-order level of analysis, namely when estimating societal change. This observation raises questions about how to improve both the accuracy and the utility of psychological scientists' expert judgments. At the least, minimal guidelines for assessing confidence in, and interpretation of, expert judgment may be beneficial (IJzerman et al., 2020). For example, in the present work, both expert and lay participants tended to predict more negative outcomes than actually unfolded, consistent with past research showing a negativity bias in predictions about the collective future (Shrikanth et al., 2018; Yamashiro & Roediger III, 2019). Keeping this tendency in mind might help both experts and policymakers correct for such biases when considering such predictions.

It is also worth noting that psychological scientists reported less confidence that their predictions would come to pass. Thus, even if they are similar to those of laypeople, there may be benefits to considering psychological scientists' predictions in aggregate if they are weighted in some way to take their level of uncertainty into account, or if such expressions of uncertainty lead to more measured or contingent policy planning.

More broadly, the present findings suggest considerable room for improvement in psychological scientists' ability to predict real-world trends. Indeed, our work, along with prior endeavors, such as the Good Judgment Project, suggests that forecasting the future is difficult. To the extent that psychologists want to make predictions for such events – which our work shows they seem willing to do and are expected to do so well by the public – then it may be advantageous for psychological scientists to learn strategies that improve forecasting accuracy at both the group (Morgan, 2014) and individual level (Grossmann et al., 2021; Mellers et al., 2019).

References

Ackerman, J. M., Tybur, J. M., & Blackwell, A. D. (2021). What role does pathogen-avoidance psychology play in pandemics? *Trends in Cognitive Sciences*, *25*(3), 177–186. https://doi.org/10.1016/j.tics.2020.11.008

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of Clinical Judgment Project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Armstrong, J. S. (1985). *Long-range forecasting: From crystal ball to computer.* John Wiley & Sons.

Artazcoz, L., Benach, J., Borrell, C., & Cortès, I. (2004). Unemployment and mental health: Understanding the interactions among gender, family roles, and social Class. *American Journal of Public Health*, *94*(1), 82–88. https://doi.org/10.2105/AJPH.94.1.82

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*, *395*(10227), 912–920. https://doi.org/10.1016/S0140-6736(20)30460-8

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Faulkner, J., Schaller, M., Park, J. H., & Duncan, L. A. (2004). Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes. *Group Processes and Intergroup Relations*, *7*(4), 333–353. https://doi.org/10.1177/1368430204046142

Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1640), 1279–1285. https://doi.org/10.1098/rspb.2008.0094

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*(1), 34–45. https://doi.org/10.1037/0096-3445.120.1.34

Gardner, R. M., & Dalsing, S. (1986). Misconceptions about psychology among college students. *Teaching of Psychology*, *13*(1), 32–34. https://doi.org/10.1207/s15328023top1301_9

Gardner, R. M., & Hund, R. M. (1983). Misconceptions of psychology among academicians. *Teaching of Psychology*, *10*(1), 20–22. https://doi.org/10.1207/s15328023top1001_5

Griskevicius, V., Tybur, J. M., Delton, A. W., & Robertson, T. E. (2011). The influence of mortality and socioeconomic status on risk and delayed rewards: A life history theory approach. *Journal of Personality and Social Psychology*, *100*(6), 1015–1026. https://doi.org/10.1037/a0022403

Grossmann, I., Dorfman, A., Oakes, H., Santos, H. C., Vohs, K. D., & Scholer, A. (2021). Training for wisdom: The distanced self-reflection diary method. *Psychological Science*, *32*(3), 381–394. https://doi.org/10.1177/09567976209691

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hawkley, L. C., & Cacioppo, J. T. (2010). Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of Behavioral Medicine*, *40*(2), 218–227. https://doi.org/10.1007/s12160-010-9210-8

Heider, F. (1958). The naive analysis of action. In F. Heider (Ed.), *The Psychology of Interpersonal Relationships* (pp. 79–124). John Wiley & Sons Inc.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Horn, H. S. (1978). Optimal tactics of reproduction and life-history. *Behavioural Ecology: An Evolutionary Approach*, 411–429.

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., Ivory, J. D., & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*(11), 1092–1094. https://doi.org/10.1038/s41562-020-00990-w

Kelly, G. A. (1955). *The psychology of personal constructs. Volume 1: A theory of personality*. WW Norton and Company.

Kendler, K. S., Karkowski, L. M., & Prescott, C. A. (1999). Causal relationship between stressful life events and the onset of major depression. *American Journal of Psychiatry*, *156*(6), 837–841. https://doi.org/10.1176/ajp.156.6.837

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. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Tetlock, P., & 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

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111

Murray, D. R., Trudeau, R., & Schaller, M. (2011). On the origins of cultural differences in conformity: Four tests of the pathogen prevalence hypothesis. *Personality and Social Psychology Bulletin*, *37*(3), 318–329. https://doi.org/10.1177/0146167210394451

Na, J., Grossmann, I., Varnum, M. E. W., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences*, *107*(14), 6192–6197. https://doi.org/10.1073/pnas.1001911107

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science*, *238*(625–631). https://doi.org/10.1126/science.3672116

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology*, *127*(5), 893–904. https://doi.org/10.1093/oxfordjournals.aje.a114892

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16(2)*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Ruggeri, K., Stock, F., Haslam, S. A., Capraro, V., Boggio, P., Ellemers, N., Cichocka, A., Douglas, K., Rand, D. G., & Cikara, M. (2022). *Evaluating expectations from social and*

behavioral science about COVID-19 and lessons for the next pandemic. https://doi.org/10.31234/osf.io/58udn

Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, *20*(2), 99–103. https://doi.org/10.1177/0963721411402596

Shrikanth, S., Szpunar, P. M., & Szpunar, K. K. (2018). Staying positive in a dystopian future: A novel dissociation between personal and collective cognition. *Journal of Experimental Psychology: General*, *147*(8), 1200–1210. https://doi.org/10.1037/xge0000421

Taylor, A. K., & Kowalski, P. (2004). Naïve psychological science: The prevalence, strength, and sources of misconceptions. *Psychological Record*, *54*(1), 15–25. https://doi.org/10.1007/BF03395459

Tetlock, P. E. (2005). *Expert political judgement: How good is it?* Princeton University Press.

Tybur, J. M., Inbar, Y., Aarøe, L., Barclay, P., Barlowe, F. K., De Barra, M., Beckerh, D. V., Borovoi, L., Choi, I., Choik, J. A., Consedine, N. S., Conway, A., Conway, J. R., Conway, P., Adoric, V. C., Demirci, D. E., Fernández, A. M., Ferreirat, D. C. S., Ishii, K., … Žezelj, I. (2016). Parasite stress and pathogen avoidance relate to distinct dimensions of political ideology across 30 nations. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(44), 12408–12413. https://doi.org/10.1073/pnas.1607398113

van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., … Willer, R. (2020).

Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*(1), 1–5. https://doi.org/10.1038/s41562-016-0015

West, R., Michie, S., Rubin, G. J., & Amlôt, R. (2020). Applying principles of behaviour change to reduce SARS-CoV-2 transmission. *Nature Human Behaviour*, *4*(5), 451–459. https://doi.org/10.1038/s41562-020-0887-9

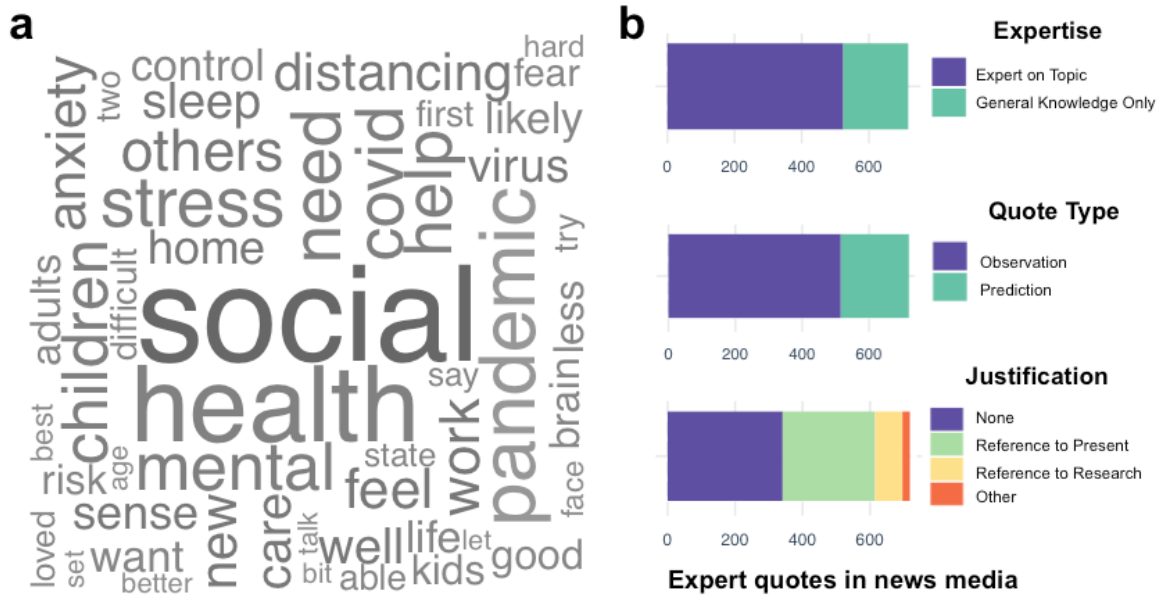
Yamashiro, J. K., & Roediger III, H. L. (2019). How we have fallen: Implicit trajectories in collective temporal thought. *Memory*, *27*(8), 1158–1166. https://doi.org/10.1080/09658211.2019.1635161

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Figures

Figure 1.

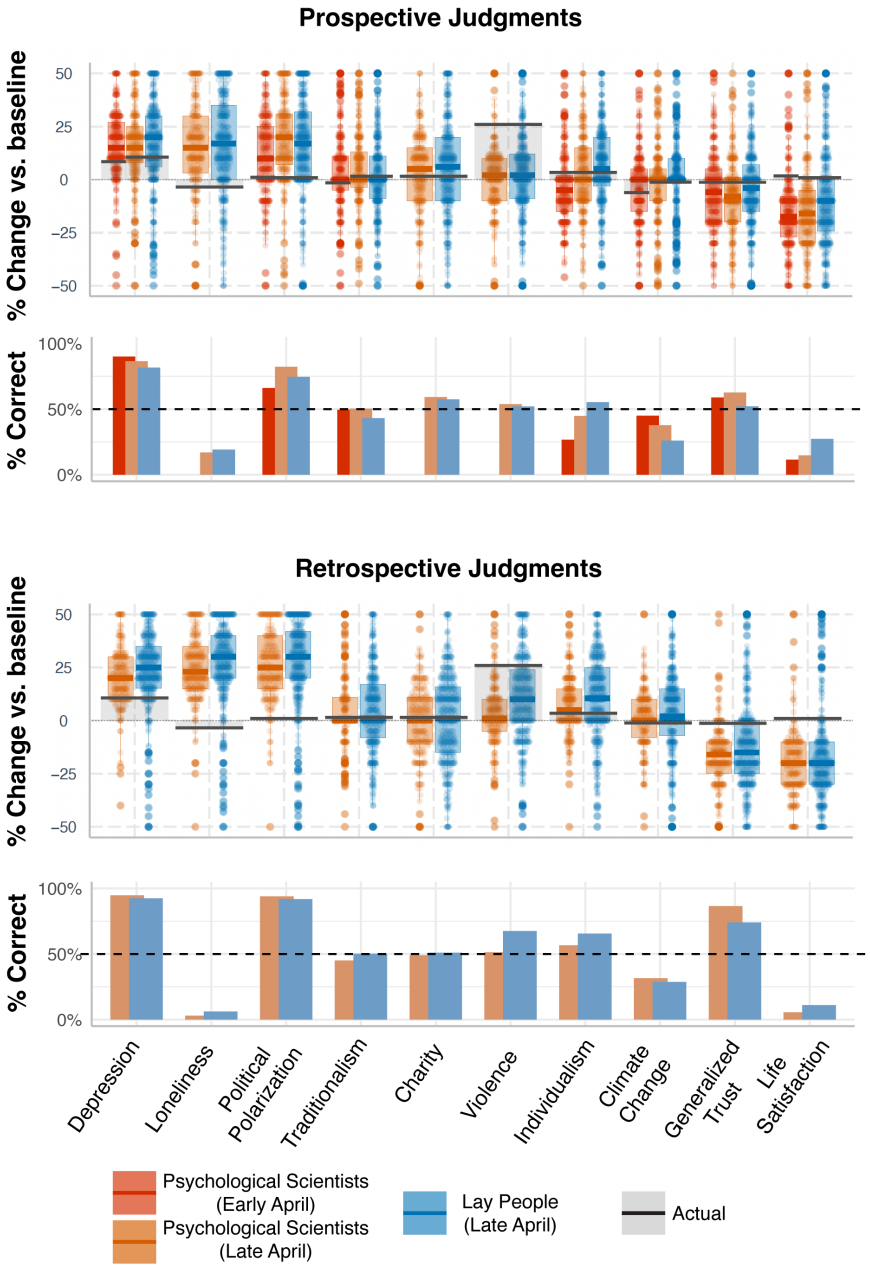
Psychological scientists in the media.



Note. a) Analysis of the frequency of different words in media interviews with experts shows that they commented on a number of topics, including health, mental well-being, stress, and social relationships. b) Analysis of these interviews also suggests that psychologists frequently spoke outside their domain-specific topic of expertise, frequently made predictions about future outcomes of the pandemic, and that quoted justifications typically involved intuitive reasoning rather than reference to specific research findings.

Figure 2.

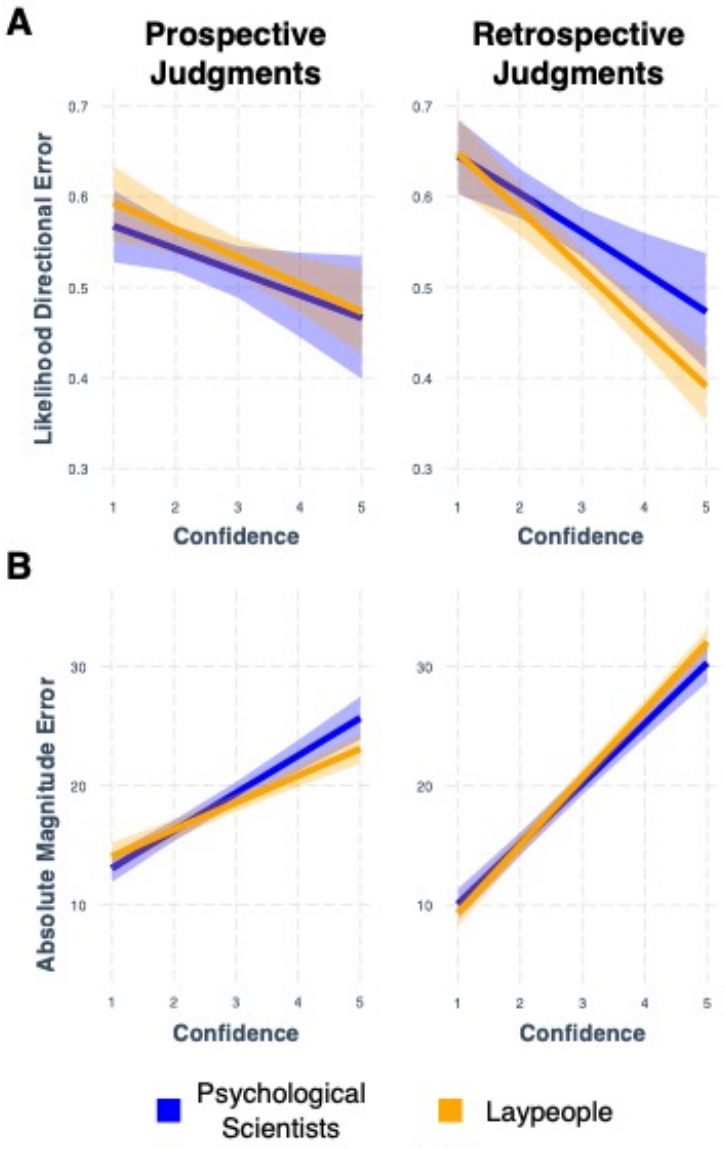
The accuracy of prospective (April and May 2020) and retrospective (October/November 2020) judgments of societal change.



Note. Predictions, along with objective markers for ten available domains, are displayed for prospective (top) and retrospective (bottom) judgments in psychological scientists and laypeople. Box-plots show median and 25/75% confidence intervals. Accuracy (measured as directionally correct predictions) is displayed just below predictions for prospective and retrospective judgments. Note: Prospective data includes two separate samples of psychological scientists surveyed in late March/early April and late April/early May). We thus display objective benchmark data separately for the two time periods where it is available. Retrospective data included a single sample of psychologists and laypeople collected in late October/early November.

Figure 3.

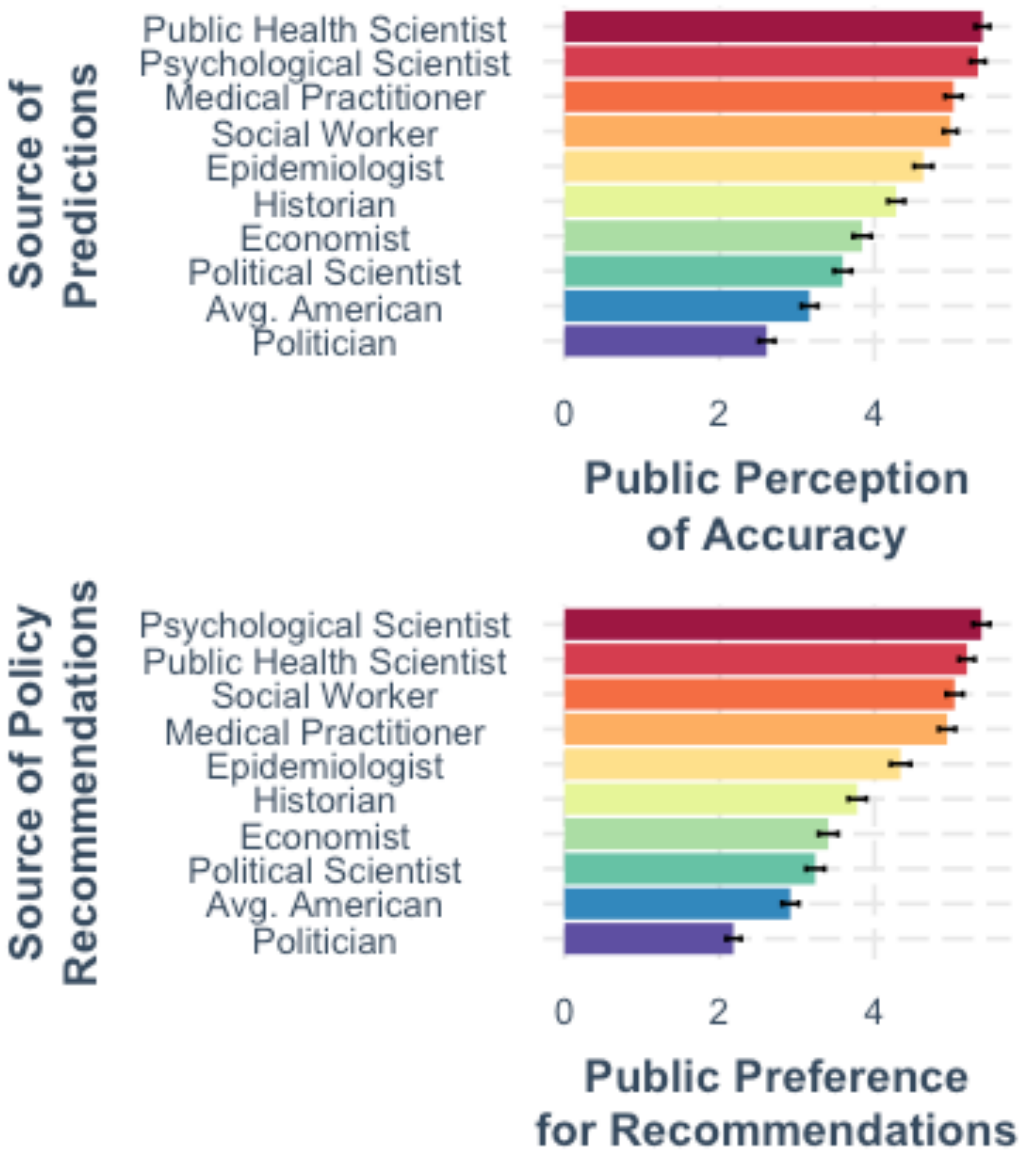
Confidence and its association with inaccuracy.



Note. Panel A: Relationship between confidence and likelihood of directional inaccuracy across domains, as a function of group (lay individuals or psychological scientists) and type of judgment (prospective predictions made in April/May of 2020, retrospective estimates made in October/November of 2020). Panel B: Relationship between confidence and average absolute error (i.e., |estimated change – actual change|). In both panels, lines and error bars display the mixed effects regression estimated line of best fit.

Figure 4.

Comparative perceptions of psychological scientists by the public.



Note. The lay public generally expects psychological scientists to be among the most accurate in predicting consequences of the pandemic for mental and social well-being, and prefers to obtain policy recommendations for dealing with these issues from psychological scientists, rather than experts in other topics like medicine, economics, or political science.

Supplementary Information

for

On the accuracy, media representation, and public perception of psychological scientists' judgments of societal change

Contents

Supplemental Methods

Study 1

Coding of media interviews

Three trained coders read the extracted text for each separate topic that a psychological scientist commented on in texts published between March 15 and May 15, 2020. Texts were coded by two coders, and any disagreements were resolved by the third coder, in consultation with the first and the last authors. Coders provided three sets of rating for each topic/item:

1) Match between scientist's domain of expertise and interview topic (inter-rater agreement = 80%, Cohen's κ= .54). This was coded by accessing the scientist's personal and professional websites to identify the domain expertise, and then judging the match between that expertise and the specific topic being commented on.
2) The type of justification given (inter-rater agreement = 79%, Cohen's κ = .66). This was given one of five possible codes: no justification; reference to current events (observing that [x] is true of the present, therefore [y] must be true); reference to historical analogies (observing that [x] was true in the past, therefore [y] must be true now); reference to research (observing that research shows that [x] is true, therefore [y]); other.
3) Whether the judgment was stated with certain or uncertain language (inter-rater agreement = 85%, Cohen's κ = .70): coded by observing whether scientists used words like "will" (certain) or "may", "might" (uncertain).

Studies 2-3

Sample size and power

For Study 2a we did not set an *a priori* criterion for sample size recruitment, aiming to recruit the largest number of participants we could in the available time. A power analysis suggested that regression analyses would have 90% power to determine small effects (f^2 = .1) of educational status with a sample size of 130. Our sample sizes more than doubled this number.

In Study 3c and 4b, to match the Study 3b and 4a samples, we targeted a nationally stratified sample of 400 lay individuals. We also aimed to recruit as many psychological scientists as we could within the 2-week availability period of the survey, targeting between 250-400 psychological scientists. While the lay sample recruitment gave us full control over the sample size, the psychological scientist sample recruitment did not. We pre-registered a stopping rule, continuing to recruit psychological scientists for 2-weeks after advertising the survey via the same venues as Study 3a, and terminating data collection after this period. This procedure ensured a roughly homogeneous time period for obtaining retrospective reports. As in Study 3a, power analyses for a small effect size (d = .2, α = .05/ β = .20) of a two-sample t-test suggested that the sample sizes obtained were adequate, especially when considering the within-subject component of our design (each participating providing 15 ratings, one for each domain)

Exploratory measures (Study 3 only)

We sought to understand the factors contributing to participants' reasoning while estimating societal change. As outlined in pre-registered exploratory analyses, we focused on the moderating role of two social-cognitive variables: a) vividness/concreteness of personal experience with a domain; b) extent of news exposure to a domain. We sought to explore whether domains in which participants report concrete visualization of personal experiences or specific news (coded as present/absent) or brought to mind vivid memories that affected them personally (also coded as present/absent) would show more extreme average retrospective estimates. In other words, domains in which retrospective estimates were positive should show more positive estimates if concrete visualizations also accompanied them. On average, domains with negative retrospective estimates should show more negative estimates if concrete visualizations accompanied them.

We also sought to explore how concreteness and news exposure contributes to alignment of retrospective estimates in Study 4 and prospective estimates in Study 3. On the one hand, construal level theory of psychological distance (1) would predict that more concrete representation of estimates would lead to greater divergence from abstract prospective estimates. On the other hand, bringing concrete events to one's mind may result in greater use of heuristics (2), biasing one's retrospective estimates toward the extreme end.

To assess *concreteness* in reasoning about social change, after completing assessments of change and confidence for each domain, participants were presented with a prompt:

> As you were reflecting on possible changes in different domains of life in the last half year, for which of the following domains did you consider news reports or specific events that occurred in the last 6 months?

To assess *vividness of memories* in reasoning about social change, participants were presented with a prompt:

> For which of these domains did you bring to mind experiences that have occurred in the past 6 months, that affected you personally, and for which you have very vivid memories?

For each question, participants were presented with check-box options, with domains presented in the same randomized order as presented earlier.

Confidence ratings (Studies 2-3)

In Study 2a, we asked psychological scientists to provide an "estimate of the probability for this forecast being true (i.e., what is the likelihood that your estimate falls within 5% of the true value)?," with responses on a scale from 0 to 100. As several psychological scientist participants pointed out to us, this question was not well understood. Therefore, we a priori chose not to analyze this initial question.

In Studies 2b-3, we modified the question to a simpler question: “How confident are you in your prediction [Study 2]/estimate [Study 3]?’ We recorded responses on a 5-point scale, with exact anchor points: 1 = “not at all”, 2 = “slightly”, 3 = “moderately”, 4 = “highly”, 5 = “extremely”.

Demographics (Studies 2-3)
Across Studies 2-3, participants reported organizational affiliation, organization size, ethnicity, annual total household income, political beliefs, gender, and age. We assessed organizational affiliation on a four-point scale. We assessed organizational size on a six-point scale. Participants selected their ethnicity from one of nine categories. Then, they indicated their total annual household income. See Table S1 for category labels.

We also measured political beliefs using a 7-point scale: 1 = Progressive, 4 = Neutral, 7 = Conservative and gender with these three choices: 1 = Woman, 2 = Man, 3 = Non-binary. Participants provided their biological age by typing a number into a textbox.

In addition to basic demographic information, we asked several questions that were only applicable to psychological scientists: country of origin, academic position, and field of research, additional fields of research and areas of expertise. Participants typed their primary country of residence into a textbox and indicated their current position by selecting one of seven options: 1 = tenured faculty, 2 = nontenured faculty, 3 = adjunct professor, 4 = postdoc, 5 = graduate student, 6 = research scientist, 7 = other. They then indicated their main field of research by selection one from the following list: 1 = Psychology, 2 = Neuroscience, 3 = Medicine, 4 = Sociology, 5 = Political Science, 6 = Economics, 7 = Epidemiology, 8 = Biology, 9 = Computer science, 10 = Other.” In Studies 2-3 we asked participants to type in any additional fields of research they engaged in and to list any domain-relevant areas of expertise (e.g., prejudice, mental health, etc.). In Study 3, the open-ended domain-relevant area of expertise question was replaced with a 15-item multi-selection list where psychological scientists selected all domains they believed themselves to have expertise in.

Quantifying domain-specific expertise (Studies 2-3)
To examine whether domain-specific expertise influences forecasts and/or retrospective accuracy, in Studies 2a,b and 3a we examined psychological scientists’ self-reported research areas, quantifying them in terms of applicability for each of the forecasted domains. In Studies 2a and b, participants provided open-ended responses regarding their domain of expertise. Third and fourth authors independently categorized each of the listed research areas and subjects of study. Two coders used a grounded and iterative approach with input from the authors to code domain of expertise. First, coders decided what category of behavioral science this expertise fell into: social/personality psychology, cognitive psychology and neuroscience, clinical psychology, developmental, or other. Second, coders decided which domain of change each participant may have an expertise in. For example, a participant who said their expertise was in “prejudice” would be coded as an expert in “social/personality

psychology," specifically with expertise in "implicit prejudice" and "explicit prejudice" among the domains assessed by this study. If participants mentioned more than one area of expertise, coders were instructed to select as many domains as applied. Inter-rater reliability was high (90% agreement). Disagreements were minor (< 5%) and resolved via discussion with the senior author.

To assess domain-specific expertise in Study 3, participants indicated in a check-box survey at the end of the study whether they had received graduate training/education (i.e., taking psychology classes or researching these or related topics) in any of the fifteen domains for which they provided estimates.

Directional Accuracy Analyses: Regression Details

In the main text, we reported accuracy results using mixed-effects regression. Here we give greater detail about each of these regressions. Further details can be found in Supplemental Tables 8-10.

For the regression examing overall accuracy of prospective predictions made by psychological scientists (combining Studies 2a and 2b due to lack of a significant difference between them), we computed a mixed-effects logistic regression using the *glmer* function in *R* with the following fixed and random components:

[GLM 1] Accuracy (0/1) ~ 1 + (1|Subject)

We used the intercept term of this model to estimate the overall accuracy for each participant compared to random chance (50%). The results of this regression are reported in the main text.

To compare academic vs. lay predictions, we computed a regression with group (psychological scientists vs. lay individuals) as a fixed effect for participants in Study 2a, b and c only. We included Domain as a fixed effect in this model to account for heterogeneity across domains:

[GLM 2] Accuracy (0/1) ~ Group + Domain + (1|Subject)

To determine the significance of the difference between groups, we computed a chi-square likelihood difference test to compare this model to the null model with only an intercept and the effect of Domain, but no effect of group.

Similarly, to determine whether retrospective predictions made by psychological scientists were significantly more or less accurate than prospective predictions, we computed a mixed-effects logistic regression of the following form, using only psychological scientists from Studies 2a and b and Study 3a:

[GLM 3] Accuracy (0/1) ~ JudgmentType + Domain + (1|Subject)

We compared this model to a null model with no fixed effect of prospective/retrospective judgment type.

Because this model suggested that psychological scientists' accuracy improved slightly from prospective to retrospective reports, we computed an additional model including laypeople's prospective and restrospective reports to determine whether this improvement was significantly larger than for the average American, using the following model:

[GLM 4] Accuracy (0/1) ~ JudgmentType*Group + Domain + (1|Subject)

We then computed a chi-square likelihood difference test to compare this model to the null model with no interaction between JudgmentType and Group:

[GLM 5] Accuracy (0/1) ~ JudgmentType + Group + Domain + (1|Subject)

Note that, in all models reported in the main body of the paper, we included only a random intercept term for participant. We did not include a random intercept term for domain because we only assessed 10 domains, and these domains were not simply random realizations from a set of almost infinite possibilities for dimensions along which societal change may unfold. Rather, these domains represented a targeted number of theoretically and pragmatically motivated topics. For these specific domains, psychological theory made clear predictions about responses in the face of the pandemic. Moreover, these were domains for which we had a reasonable expectation about availability of ground truth data.

However, for completeness, and to assess the robustness of effects, we also computed variants of GLMs 1, 2, 3, and 4 in which Domains was specified as a random instead of fixed effect. For example, to compute the accuracy of prospective reports, we computed the following regression:

[GLM 1b] Accuracy (0/1) ~ 1 + (1|Subject) + (1|Domain)

This model yielded nearly identical conclusions as the results reported in the main text, with an average individual accuracy of 48.7% [31.5 65.9], a value that was not significantly different from chance, z = -.15, p = .88. Similarly identical conclusions were obtained for GLMs 2, 3, and 4. Where results were not significant and Bayes Factors favored the null hypothesis that psychological scientists' accuracy was no different than lay individuals when including Domain as a fixed effect, results were also non-significant and Bayes Factors favored the null hypothesis even more strongly. Similarly, in the one case where we observed significant differences (i.e., an improvement from prospective to retrospective reports), we found nearly identical patterns with random effects models.

Thus, although our conclusions must generally be tempered by the limited number of domains in which we were able to assess accuracy, we think they are generally robust to different model specifications.

Accuracy Benchmarks

To gauge how much each dimension changed between April/May 2020 and October 2020 at a national level we searched for representative surveys that tracked constructs of interest over time. We were able to identify 16 such surveys covering 10 domains (four options for depression, three for life satisfaction, two for loneliness and political polarization, and one for the rest) from 11 sources.

Life satisfaction

Our primary marker of life satisfaction, along with several secondary markers describe below, relied on Gallup Panel data – COVID-19 Survey. The COVID-19 web survey began fielding on March 13, 2020 with daily random samples of U.S. adults, aged 18 and older who are members of the Gallup Panel. Approximately 1,200 daily completes were collected from March 13 through April 26, 2020. From April 27 to August 16, 2020 approximately 500 daily completes were collected. Starting August 17, 2020, the survey moved from daily surveying to a survey conducted one time per month over a two-week field period (typically the last two weeks of the month). The Gallup Panel is a probability-based, nationally representative panel of U.S. adults. Members are randomly selected using random-digit-dial phone interviews that cover landline and cellphones and address-based sampling methods.

Gallup weights the obtained samples each day to adjust for the probability of selection and to correct for nonresponse bias. Nonresponse adjustments are made by adjusting the sample to match the national demographics of gender, age, race, Hispanic ethnicity, education and region.
Demographic weighting targets are based on the most recent Current Population Survey figures
for the aged-18-and-older U.S. population.

As a benchmark for life satisfaction change, we used the same time period as estimated by participants in our forecasting Study 2b,c (April 23 – May 5, 2020) and retrospective Study 3 (October 14 – October 26, 2020). Gallup panel provided the closest match in the definition of life satisfaction provided to our Study 2-3 participants, as it was assessed with classic Cantril ladder question “Please imagine a ladder with steps numbered from 0 at the bottom to ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?” The response scale ranged from 10 – best possible to 0 – worst possible. We obtained average estimates across each time point. Subsequently, we calculated

percentage change between April/May and October estimates as our primary marker of % societal change in life satisfaction.

Additional markers of life satisfaction

Twitter-based estimates
We also considered Twitter based estimates, because work suggests that national estimates obtained via social media language can reliably track subjective well-being (3). For each month, we used previously validated predictive models of well-being, as measured by life satisfaction scales (4). 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 tend to follow nationally representative polls (3). We applied the respective models to Twitter data from late April/early May 2020 and late October 2020 to obtain estimates of life satisfaction via language on social media. Estimates obtained this way were only 1% off from the Gallup estimates we selected and vastly different from the forecasted and retrospective estimates provided by our participants.

ICL/YouGov
Imperial College London partnered with YouGov to track behaviors over the period of coronavirus pandemic, and included measures of life satisfaction over time (https://github.com/YouGov-Data/covid-19-tracker). The Imperial College London YouGov Covid 19 Behaviour Tracker is a multinational COVID tracker of behavior, which included surveys for life satisfaction in the US. Unfortunately, the surveys for April and September were much smaller in scope compared to Gallup Polls and the US-based survey did not include estimates for October. Thus, we focused our analyses on the Gallup data described above.

Loneliness
We used USC's Center for Economic and Social Research (CESR) Understanding America Study (UAS) – a probability-based Internet panel. Since April, 2020, the Understanding America Study included a coronavirus tracking poll. Each panel member was randomized to respond on a pre-assigned day of the week, distributed so that a full sample is invited to participate over a 14-day period. Respondents have 14 days to complete the survey but receive an extra monetary incentive for completing the survey on the day they are invited to participate. Data for the full sample is thus final after a 28-day period. Approximately 7100 adult residents of the U.S. have been participating in the ongoing surveys; roughly 550 per day. Full information on the survey, including full methodology, is available at https://covid19pulse.usc.edu.

The survey included the following question: “In the past 7 days, how often have you felt lonely?”, with response options 1 = “Not at all or less than 1 day”, 2 – “1-2 days”, 3 – “3-4 days”, 4 – “5-7 days”. We used weighted responses for this question for each of two time periods roughly corresponding to the time windows we gathered prospective

survey data (Study 2) – April 15-May 15 and retrospective data (Study 3) – September 15-October 15.[1] We subsequently calculated % change between these average points.

We chose the Understanding America Study marker of loneliness because it had a more balanced sample size across both time windows compared to other metrics (See Table S3) and because it more closely matches the definition provided to participants.

Additional marker of loneliness

We used Gallup Panel data as a secondary marker of loneliness, assessed with the question "Did you experience the following feelings during A LOT OF THE DAY yesterday?" Response options included a range of feelings, from enjoyment, and happiness, to worry and loneliness. We examined weighted % of participants reported experiencing loneliness yesterday as a secondary marker of loneliness. We used the same time period as estimated by participants in our forecasting Study 2 (April 23 – May 5, 2020) and retrospective Study 3 (October 14 – October 26, 2020). However, the Gallup panel score for loneliness was not an ideal match for the definition of loneliness provided to our Study 2-3 participants (see Table S2 for wording), because Gallup data explicitly focused on the feeling of loneliness from the previous day (presence/absence) rather than its magnitude (e.g., "A lot of times I feel lonely," "I often feel left out of things"). Thus, we focus our primary analyses on the Understanding America Study described above. Estimates for societal change using Gallup Panel loneliness were very similar to the Understanding America Study estimates (see Table S3).

Depression

Our primary source for rate of depression came from the US Centers for Disease Control and Prevention Household Pulse Survey. The Household Pulse Survey (HPS) was launched on April 23, 2020 as a joint effort between National Center for Health Statistics (NCHS) and Census Bureau to monitor changes in mental health throughout the COVID-19 pandemic. It involved a 20-minute survey, aiming to assess frequency of anxiety and depression symptoms. The questions were modified versions of the two-item Patient Health Questionnaire (PHQ-2) and the two-item Generalized Anxiety Disorder (GAD-2) scale on the Household Pulse Survey, collecting information on symptoms over the last 7 days. Full information on the survey and methodology can be found online at https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm. We focused on weighted depression scores from this survey, targeting the dates most closely matching our forecasted and retrospective Studies 2-3: April 23-May 5, 2020 and October 14 – October 26, 2020.[2] We chose the CDC Household Pulse Survey, because it included the most extensive survey of clinical measures of depression and best matched questions we asked participants in Studies 2-3.

---

[1] Weights were used to account for non-response bias and discrepancies between the sample and population on key demographic dimensions (e.g., gender, race, ethnicity, age, education, and Census regions).

[2] The Household Pulse Survey estimates were weighted to adjust for nonresponse and discrepancies between the survey and population for age, sex, race and ethnicity, and educational attainment.

Additional markers of depression

Gallup Panel
We used Gallup Panel data as a secondary marker of depression, assessed with a question “Did you experience the following feelings during A LOT OF THE DAY yesterday?” Response options included a range of feelings, from enjoyment, and happiness, to worry and depression. We examined weighted % of participants who reported experiencing depression yesterday as a secondary marker of depression. We used the same time period as estimated by participants in our forecasting Study 2 (April 23 – May 5, 2020) and retrospective Study 3 (October 14 – October 26, 2020). Gallup panel score for depression was not an ideal match for the definition of depression provided to our Study 2-3 participants (see Table S2), because Gallup data explicitly focused on the feeling of depression from the previous day (presence/absence) rather than a clinical definition of depression we provided to participants in the forecasting and retrospective studies (characterized by feeling sad, losing interest in activities once enjoyed, and a loss of energy over a prolonged period of time, and is measured by agreement with statements like “I am sad all the time and I can't snap out of it”).

Understanding America Study
We used USC's *Understanding America Study* described above as a third benchmark of depression. Similar to loneliness, it concerned a response to two questions: “Over the past 14 days, how often have you felt feeling down, depressed or hopeless?” and “Over the past 14 days, how often have you felt little interest or pleasure in doing things?” with response options 1 = “Not at all”, 2 – “Several days”, 3 – “More than half the days”, and 4 – “Nearly every day”. We averaged the two items and then computed weighted means using poststratification weights. The responses for this question for each of two time periods roughly correspond to the time windows we gathered prospective survey data (Study 2b) – April 15-May 15 and retrospective data (Study 3) – September 15-October 15. We subsequently calculated % change between these average points. We chose not to use this index as a primary marker because it is smaller and less representative of clinical depression compared to the HPS above.

Affective polarization
Our primary marker of affective polarization, as well as several other indices below, relied on nationally representative data from Nationscape (5). Nationscape is a survey that conducted 500,000 interviews of Americans from July 2019 through December 2020, covering the 2020 campaign and election. The survey was in the field starting July 10, 2019, and included interviews with roughly 6,250 people per week. Nationscape samples were provided by Lucid, a market research platform that runs an online exchange for survey respondents. The samples drawn from this exchange match a set of demographic quotas on age, gender, ethnicity, region, income, and education. Respondents were sent from Lucid directly to survey software operated by the Nationscape team. All respondents took the survey online and completed an attention check before taking the survey. The survey was conducted in English. The Nationscape

data are weighted to be representative of the American population. The weights are generated using a simple raking technique, as there is little benefit to more complicated approaches (6). Nationscape generated a set of weights for each week's survey. The targets to which Nationscape is weighted were derived from the adult population of the 2017 American Community Survey of the U.S. Census Bureau. The one exception was the 2016 vote, which was derived from the official election results released by the Federal Election Commission. The Nationscape team weighted the data on the following factors: gender, the four major census regions, race, Hispanic ethnicity, household income, education, age, language spoken at home, nativity (U.S.- or foreign-born), 2016 presidential vote, and the urban-rural mix of the respondent's ZIP code. Data were also weighted on the following interactions: Hispanic ethnicity by language spoken at home, education by gender, gender by race, race by Hispanic origin, race by education, and Hispanic origin by education. More information on the survey can be found at www.voterstudygroup.org.

Following Boxell, Conway, Druckman, and Gentzkow (7), we conceptualized affective polarization via responses to the question stating, "Here are the names of some groups that are in the news from time to time. How favorable is your impression of each group or haven't you heard enough to say?" and containing responses for "Very favorable," "Somewhat favorable," "Somewhat unfavorable," "Very unfavorable," and "Haven't heard enough." The survey then goes on to ask about the groups: "Republicans" and "Democrats." We code "Very unfavorable" through "Very favorable" from 0 to 3 respectively and we exclude respondents with other responses. Affective polarization at time t was then defined as:

$$\pi_t = \frac{1}{\bar{N}_t} \sum_{i \in N_t} w_i \left(A_i^{P(i)} - A_i^{\sim P(i)} \right)$$

where N_t is the set of respondents in period t identifying with either the Republican or Democratic party who have valid affect responses for both parties, w_i is the survey weight, and $N_t = \sum_{i2Nt} w_i$. Affective polarization measures average feelings towards one's own party minus average feelings towards the opposing party.

We use the periods overlapping with the time we gathered prospective data (Study 2) and retrospective data (Study 3)—April 23-May 6, 2020 and October 1-October 28, 2020. We restricted survey observations to respondents that give a valid state. We estimated change by examining % change of the October 2020 score relative to the April 2020 scores.

Additional marker of affective polarization

We considered Gallup poll data of presidential approval ratings by party identification as an alternative marker (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. We did not pursue this marker for primary analyses because it does not fully

capture affective attitudes toward members of the other party and hence is not fully in sync with the definition provided to our participants. This said, respective estimates of societal change from this marker were within 1.5% of the estimate of the affective polarization marker from Nationscape.

Individualism

We used the COVID-19 attitudes survey (Neuberg, Varnum, Becker, Ko, Pick, & Wormley, 2020) to estimate societal change in individualism. Using the Prolific survey collection platform, researchers collected two nationally representative samples of US residents to examine the effects of the coronavirus pandemic on a variety of behaviors. The dataset included information gathered at two time points close to the time we gathered prospective estimates (Study 2) and retrospective estimates (Study 3). Upon exclusion of incomplete responses or returned submissions, relevant waves from this project included substantial number of participants surveyed on April 22, 2020 (N = 1,510) and on September 23, 2020 (N = 805).

To assess individualism, participants rated their agreement with three statements (1 = Strongly Disagree, 9 = Strongly Agree) “It is better for me to follow my own ideas than to follow those of anyone else,” “I enjoy being unique and different from others in many respects,” and “My personal achievements and accomplishments are very important to who I am.” Items came from the established individualism scale (8). Given the multi-item nature of the measure, we first inspected measurement invariance by comparing inter-item zero-order correlations at each time point. These preliminary results indicated a variable degree of inter-item association at time 1, .22< rs ≤ .35, and at time 2, .16< rs ≤ .36. Therefore, we selected two items with highest and largely comparable correlations at both time points, which concerned the first and the second items (time 1: r = .35; time 1: r = .36). We averaged these items, prior to performing a weighting procedure to ensure the responses represent US population. Like with Nationscape raking procedure, we weighted responses for race, gender, education, age, and political orientation, at each time point. Subsequently, we calculated percentage difference between April and late September estimates as a marker of societal change in individualism.

Generalized Trust

We used the same database as for individualism described above (Neuberg et al., 2020). Researchers measured generalized trust with a single item measure from prior research (9):“Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?” with 1 indicating “You can’t be too careful” and 9 indicating “Most people can be trusted.” We applied the raking weighting procedure as described above, and calculated percentage difference between April and late September estimates as a marker of societal change in generalized trust.

Traditionalism

Using the *Nationscape* data described above, we examined weighted % of people choosing the response option "The government should promote traditional family values in our society" instead of the option "The government should not promote traditional family values in our society." We calculated the difference in percent participants agreeing to this question in surveys conducted in April 2020 (April 23-May 6 – same period as Study 2) and October 2020 (October 1-28, 2020 – same period as Study 3) as a marker of change in endorsement of traditionalist values.

Additional marker of traditionalism

To obtain another marker of traditionalism, we used data from an on-going project on societal attitudes during COVID-19 described above (Neuberg et al., 2020).

To assess traditionalism, participants rated their agreement with three statements (1 = Strongly Disagree, 9 = Strongly Agree): "Traditions interfere with progress,"* "People should respect social norms," and "Traditions are the foundation of a healthy society and should be respected" (* = reverse coded). Items came from the established traditionalism scale (10). Given the multi-item nature of the measure, we first inspected measurement invariance by comparing inter-item zero-order correlations at each time point. Preliminary results indicated a variable degree of association at time 1, $.33 < r\text{s} \leq .58$, and at time 2, $.42 < r\text{s} \leq .63$. Therefore, we selected two items with the highest and largely comparable correlations at both time points, which concerned the first and last items (time 1: $r = .55$; time 1: $r = .66$). We averaged these items, prior to performing a weighting procedure to ensure the responses represent US population. Like with Nationscape raking, we weighted responses for race, gender, education, age, and political orientation, at each time point. Subsequently, we calculated percentage difference between April and late September estimates. Because the time frame for this estimate was a month shorter than for Nationscape data, we chose to treat this estimate as a secondary benchmark. We note that the estimates of societal change in traditionalism were very similar across both primary and secondary markers (within 5% change).

Violence

To assess changes in violent crime between April and October of 2020, we relied on data from the Pandemic, Social Unrest, and Crime in U.S. Cities November 2020 report, prepared for the Council of Criminal Justice (11); https://cdn.ymaws.com/counciloncj.org/resource/resmgr/covid_commission/Crime_in_US_Cities_-_October.pdf). It tracks ten types of criminal offense from January 2017 for 28 U.S. cities, spanning a population of 866,000 people. Out of these 10 only four met the definition of violent crime provided to participants in Studies 2-3, as they were specifically violent in nature: aggravated assault, homicide, gun assault, and domestic violence. We thus calculated a percentage difference score for each crime type and then created a composite by averaging all four.

Attitudes toward climate change

Using *Nationscape* data described above, we examined weighted % of people agreeing to the question "We'd like to know whether you would cap carbon emissions to combat climate change," with response options "agree," "disagree," "not sure." We calculated the difference in percent participants agreeing to this question in surveys conducted in April 2020 (April 23-May 6 – same period as Study 2) and October 2020 (October 1-28, 2020 – same period as Study 3) as a marker of attitudes toward climate change.

Charitable Giving

We obtained estimates from charitable donation data for the US collected by Giving Tuesday to estimate philanthropic sentiment (http://data.givingtuesday.org/), assessed as part of the Fundraising Effectiveness Project. The Fundraising Effectiveness Project and the Growth in Giving database (created in 2012) are administered by the Association of Fundraising Professionals. The Growth in Giving database is the world's largest public record of donation activity, with more than 204 million donation transactions, and is continuously updated by leading fundraising software thought leaders (in alphabetical order) Bloomerang, DonorPerfect, and NeonCRM. Additional partners include the 7th Day Adventists, The Biedermann Group, DataLake Nonprofit Research, and DonorTrends (a division of EveryAction). We specifically focused on the number of people in the US donating to charities in April/May and October/November, 2020.

Supplementary benchmark indices

We initially planned to include two additional markers concerning explicit and implicit prejudice toward minorities. The Project Implicit data source is not representative of the US population at large and relies on different on-line platforms through which participants are recruited. Because the topic concerned prejudice, and Black-Lives Matter protests in the summer let many outlets and diversity programs directing persons interested in learning about empathy and prejudice to the website, the representativeness of the data could be viewed as compromised. Out of an abundance of caution, we decided not to report this benchmark estimate in the main text. For the sake of transparency, we report all relevant analyses in this supplement.

This data came 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 - 60,000 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 (12, 13) and when studying cross-temporal societal shifts in U.S. attitudes (13). 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, indicating that group aggregates of the bias data from Project Implicit can reliably approximate group-level estimates (14). To correct possible non-

representativeness, we applied stratified weighting to the estimates, as described below.

Because of possible selection bias among the Project Implicit participants, we used a raking procedure similar to the one employed by Nationscape. We weighted 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.

To correct for possible variability in monthly scores due to fluctuations in sources of participant recruitment, we further applied 30% loess smoothing function across monthly estimates from 2018 through 2020, prior to calculating % change scores between April (April 1 – 30) and October (October 1 – 31), 2020. This approach allows to correct for month-specific selection biases.

Explicit prejudice
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). We calculated relative explicit bias as the difference in responses to minority and majority groups on feeling thermometers (for Asian-American and Black Americans). The sample was further restricted to include only respondents from the United States to increase shared cultural understanding of attitude categories. The sample was also restricted to include only respondents with complete demographic information on age, gender, race/ethnicity, and political ideology. After raking and smoothing, we averaged responses across both estimates for an overall measure of bias toward ethnic minorities.

Implicit prejudice
Implicit attitude scores were computed using the revised scoring algorithm of the implicit association test (IAT) (15). The IAT is a computerized task comparing reaction times to categorize paired concepts (in this case, social groups, e.g., Black American vs. European American and 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 averaged responses across both estimates for an overall measure of bias toward ethnic minorities.

Accuracy Benchmarks at 12 Months

We were also able to obtain accuracy benchmarks at 12 months for four (depression, loneliness, violence, charity) out of the 10 of the same domains we tracked at 6 months (see method above). Additionally, we acquired estimates for birth rates which were derived using data from the The Human Fertility Database (HFD). The HFD is a joint project of the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany and the Vienna Institute of Demography (VID) in Vienna, Austria, based at MPIDR. The HFD is a high quality dataset designed for making fertility comparisons across time and countries (https://www.humanfertility.org/cgi-bin/main.php). The birth rate benchmark was calculated by computing the % change in the total number of recorded births in the US in April 2020 as compared to April 2021. See Table S3 for % change estimated from each of these sources.

Study 4

Question Wording

For predictions, participants received the prompt: “Imagine that we polled groups of people (below) about how COVID-19 would affect societal changes in depression, life satisfaction, loneliness, violence and related domains in the next half a year. To what extent would you expect these groups to make accurate predictions for these social trends in your country?”

For recommendations, participants received the prompt: “Now imagine we polled each group about what they think society should do to address societal issues concerning depression, life satisfaction, loneliness, violence and related issues resulting from the pandemic. Who would you like to make recommendations for these societal issues? Please rate how much you'd prefer hearing from each group below.” For each question, participants provided responses on a 7-point scale (not at all / a little / somewhat / a moderate amount / a good deal / a lot / very much).

For ranking, participants received the prompt: “Imagine that you want to get a good idea about how the COVID-19 pandemic will affect human behavior and society in the long-term. Who would you want to ask? Pick your top three, ranking them in order from 1 (most preferred) to 3 (less preferred).” For analyses, we recoded ranking responses from “not selected” = 0 / third rank – 1 / second rank – 2 / first rank = 3.

Sample Size and Power

We targeted 200 lay individuals. Given the within-subject design, this sample size was sufficient to detect a small effect size (r = .12, α = .05/ β = .20) of a two-sample t-test suggested that the sample sizes obtained were adequate. We did not have a target for the supplementary sample of academics/policy-makers, and aimed to recruit as many participants as we could.

Supplemental Results

Studies 2-3

Deliberation check

To examine whether participants relied on intuition or spent a substantial amount of time reflecting on predictions, we examined descriptive statistics for overall study completion time among participants who completed the whole survey. In Study 2, psychological scientists typically took 11 *min* (median *Md*; mean M = 23.80; 95%*CI* [16.22, 31.38]) in total. In Study 3, they spent 14 *min* (*Md*; *M* = 20.72 95%*CI* [17.76, 23.67]) in total, whereas lay people spent 12 *min* (*Md*; *M* = 14.47; 95%*CI* [13.60, 15.33]). Consequently, participants typically spent less than a minute making predictions for each of the eleven (Study 2a) / fifteen (Study 2b-3) domains. This suggests that most of the participants’ predictions made were not the result of protracted reflection. In Study 3, we collected domain-specific times of completion, rather than simply completion time for the whole survey. By this measure, psychological scientists took on average between 10 and 20 seconds (M = 14.48; $9.05 < 95\%CI \leq 21.63$) to read relevant descriptions and subsequently answer two questions per

domain. In comparison, lay people took between 11 and 25 *seconds* per domain (M = 16.41; 10.20 < 95%CI ≤ 26.74). Thus, it appears that retrospective estimates were likewise not the result of protracted reflection, nor did psychological scientists deliberate for longer amounts of time.

Description of predictions by domain

In addition to examining accuracy, we also analyzed data from Studies 2-3 for general predictions about change and whether those changes would return to baseline within the next two years.

We begin by focusing on Study 2a, which took place at the beginning of April 2020. Figures S2-S3 display predictions for change across different domains. Psychological scientists predicted the largest changes for depression, political polarization, out-group prejudice, and life satisfaction (Figure S4). Notably, for three of the eleven domains (traditionalism, generalized trust, delay of gratification) psychological scientists' predictions for April 2022 were not statistically different from the baseline in April 2020, *ps* > .072; thus, for these domains psychological scientists predicted a full return to baseline. Psychological scientists predicted the remaining domains to remain significantly altered two years later, *ps* < .028 (see Table S4[3]).

Did psychological scientists' prospective intuitions shift over short periods of time? To assess this question, we turn to Study 3b, conducted at the beginning of May 2020. As in Study 3a, psychological scientists predicted a significant degree of societal change for each of these domains, 2.76 < *t*s ≤ 16.86, *p*s < .007 (see Table S5 and Figure S3 for comparison of April and May 2020 estimates by psychological scientists), except for delay of gratification (replicating Study 3a), p = .100. Psychological scientists predicted that traditionalism, birth rate, delay of gratification, and charitable donations would return to May 2020 baseline in two years, while for the remaining domains psychological scientists expected a significant difference two years from May 2020, *ps* < .033 (see Table S4). Notably, a comparison of estimates across Studies 3a and 3b revealed highly similar patterns of forecasts, both for temporal trends (see Figs. S2-S3), and rank-order of trends (compare Panels A-B in Figure S4), Bayes Factor (null / alternative, BF01) > 6 (see Table S5 for frequentist results). We observed only four exceptions: different linear trends for individualism, birth rate and political polarization and quadratic trends for birth rates and life satisfaction, Bayes factor < 2.

Comparing lay and academic predictions in May of 2020

We also fit a second model using the same procedure as above, except we compared lay people's and psychological scientists' forecasts in May 2020. The results indicated no significant difference in lay people's versus psychological scientists'

[3] Beyond the domains provided in a questionnaire format, analyses of open-ended responses revealed that psychological scientists identified health and well-being (mental illness, psychological and physical well-being), interconnectedness (romantic relationships, social norms), economics (economic concerns, health care attitudes), social justice (inequality, poverty), child development (education, child development), political discord and mistrust in institutions (science denialism, right-wing orientation) as key domains of pandemic-related societal change.

prediction except for explicit prejudice (linear trend) and birth rate (quadratic trend; Table S6).

Sensitivity of analyses to precise time period of benchmarks

One potential concern with the analyses reported in the main text centers around the fact that we collected predictions from two different samples of academics, one completing those predictions in late March/early April (Study 2a) and another completing those predictions in late April/early May (Study 2b). In the main text, we compared both sets of predictions to a single set of benchmarks designed to match the time period of interest for Studies 2b and 2c, as well as Studies 3a and 3b (retrospective estimates from late October/early November). These benchmarks use as a baseline the time period in late April/May and the time period in October/early November for estimating change. However, this methodological choice raises a question about whether benchmarks from the relevant time period for Study 2a (i.e., baseline in March/early April and benchmark change in late September/early October) would yield different results. In other words, would our conclusions differ depending on the exact time period against which we compared estimates of societal change?

For many of the measures we used in this study, high-quality data only began to be collected in mid-late April or early May, limiting our ability to draw firm conclusions about this issue for all domains. However, for a subset of domains (life satisfaction, affective polarization, climate change, and explicit/implicit prejudice) we had data with the necessary temporal resolution in early April of 2020 and early October of 2020. We also were able to obtain birth rate data in March and April of 2020 to compare to 12-month predictions in March and April of 2021. Notably, for all domains with the exception traditionalism, our alternative time periods showed similar changes in both direction and magnitude (see Table S3 for relevant details). Thus, it is unlikely that our conclusions depend substantially on choice of exact time period, with the caveat that this inference is based on a subset of data covering only six domains.

Alternative measures of accuracy: Absolute magnitude of accuracy

In the main text, we report on measures of accuracy in terms of direction (did participants successfully predict whether an outcome would increase or decrease?). We also sought to examine how conclusions might change with alternative definitions.

One way to test whether participants were accurate is to compare the estimated magnitude of change against observed benchmarks. We thus subjected psychological scientists' and lay people's prospective and retrospective estimates to a series of one sample t-tests with *mu* set to the accuracy benchmark level retrieved from nationally representative samples by domain, with a subsequent Benjamini-Hochberg correction. As Table S7 shows, for most domains prospective and retrospective estimates of change were significantly different from actual changes.

We also sought to assess accuracy by quantifying how many participants were accurate to within a certain percentage of the true change, using three benchmarks of decreasing stringency: i) being within 1% point of the actual estimate (bound of half a percent point on each side of the accuracy estimate); ii) being within 5% point of the actual estimate (bound of 2.5% on each side of the accuracy estimate); iii) being within

20% point of the actual estimate (bound of 10% on each side of the accuracy estimate). We compared the percentage of participants within each benchmark by estimate type (prospective / retrospective), sample type (lay / expert) and domain type.

Using this alternative metric, we again found little evidence that predictions of societal change were accurate overall, as for ten domains they were off by an average of 18%, Range = 3% - 64%, *ps* < .002. However four (out of 15) domains showed some evidence of accuracy: predictions regarding charitable giving, individualism, climate change and traditionalism were off by an average of only 1%, Range = 1% – 2%, *ps* > .195 (see Table S7 for statistical tests). Retrospective estimates of change were similarly inaccurate, *ps* < .015, except for climate change beliefs, *p* = .064 (see Table S7). However, even for climate change beliefs psychological scientists were largely inaccurate in estimating the direction of change, with only 26% estimated it correctly. Beyond these few domains, estimates were on average strikingly inaccurate compared to objective markers.

Figure S5 quantifies the percentage of accurate responses for each sample, using both strict and more liberal percentage-difference cutoffs as a measure of accuracy. Using a strict criterion (within 1% point of the estimate), in most domains, less than 2 % of each sample were accurate in their forecasts, with somewhat better estimates for traditionalism, life satisfaction, generalized trust and depression rates. Using a moderate criterion (within 5% of the estimate), for most domains, less than 10% of each sample was accurate, except for traditionalism, depression, climate change beliefs, generalized trust and charity (for scientists).

Making predictions is difficult especially when the outcomes might be influenced by conditional factors (e.g., will governments enact fiscal stimulus, will masks and social distancing be required or only encouraged, will a vaccine be developed quickly?). Thus, we also assessed retrospective judgments of the pandemic's societal effects. Perhaps psychological scientists might show greater accuracy for retrospective assessment of societal change. Our results suggest that this is not the case. Retrospective estimates showed a similar if not smaller number of accurate estimates. Even when using a liberal criterion (within 20% of the estimate), for most domains, less than 41% of each sample was accurate, and for no domain did we observe a meaningful majority of participants being accurate (60+%). Notably, Figure S5 demonstrates that numbers of accurate estimates were very similar between psychological scientists and lay people, with most differences within a negligible rate of 5% difference (with the exception of retrospective estimates of depression and climate change beliefs when using moderate and liberal criteria).

Alternative measures of accuracy: Rank order accuracy

Although scientists appear not to be accurate when assessing change in a given domain, it is possible that they are more accurate when evaluating domains in relation to each other—after all, psychological scientists often study how social phenomena or processes are associated with each other and make conditional inferences. Thus, they could be more accurate when judging the rank order of most positive to most negative societal change across domains. To address this question, for each participant we calculated the rank-order correlation ρ between their estimates and objective markers across all 10 domains. Here, ρ represents the degree of accuracy in estimated

compared to objective rank order. To assess significance, we constructed a null distribution of the expected rank-order correlation using 5000 random permutations of the observed outcomes. As Figure S6 shows, psychological scientists and lay people alike had average rank-order correlations in the range of $.05 < \rho \leq .08$. Permutation tests with random shuffling of domain labels suggest that this degree of correlation is not significantly different from chance. Rank-order accuracy also did not vary by sample (psychological scientists vs. lay people), or judgment type (prospective vs. retrospective), *p*s > .594.

Role of expertise

We were interested in the effects of *expertise level* in behavioral sciences on predictions. We operationalized expertise level by categorizing our sample into three clear-cut categories: a) tenured; b) non-tenured; c) graduate students/post-doc. Tenured faculty consists of psychological scientists who chose the "tenured faculty" option as their current position at university/college. Non-tenured faculty consists of those who chose "nontenured faculty" and "adjunct professor" as their current position. Finally, the Grad Students/Post-doc group is comprised of those who selected "graduate students" and "post-docs" as their current position respectively. We fit a 3-way MLM with expertise level, time, and domain as well as all two- and three-way interaction terms predicting participants' forecasts (Figure S7). In addition, we controlled for age, university size, affiliate organization type, gender and country of residence to rule out the possibility that demographic variability between the two samples could be responsible for observed results. Observations were nested in participants. Finally, to control for the number of comparisons we used the Benjamini-Hochberg correction. We observed no significant differences in predictions between the three groups of experts, *ps* > .058 except for two domains. Tenured faculty forecasted lower levels of change at 6 months for charity, compared to non-tenured and graduate students/post-docs. Similarly, graduate students/post-docs forecasted greater political polarization then both tenured and non-tenured faculty for all time points. A mixed-effects logistic regression with directional accuracy (1 = correct/0 = incorrect) as the dependent measure and expertise level, domain, and their interaction as fixed effects, and participant as a random effect suggested that the differences in predictions among these three groups generally favored the predictive accuracy of students over faculty (see Table S11 for accuracy rates).

In addition to expertise level, we also analyzed whether domain-specific expertise mattered. Participants listed their areas of expertise, which we sorted into one of three categories: Social/Personality Psychology, Mental Health and Other, with Other encompassing all other areas of psychology and other social and life sciences. We fit a 3-way MLM model with Time and Dimension as level 1 predictors and Academic Discipline as level 2 predictor (a factor decomposed into two dummy variables) as well as all 2 and 3-way interaction terms. The model also included the following socio-demographic covariates: age, university size, organization, gender, and country of residence. Fitted means and confidence intervals were extracted from the model and were then used in plotting (Figure S8). To test for differences between expertise categories, we conducted pairwise comparisons in *R* using *emmeans* and controlled for

number of tests using the Benjamini-Hochberg correction. The differences between areas of expertise at each time point and domain were not statistically significant, *ps* > .55.

Finally, we asked whether psychologists made different predictions from other groups of scientists. Psychology group included experts who selected either "Psychology" or "Neuroscience" as their main field of research. We combined the remaining choices into "Other disciplines." Then, we calculated differences between the two conditions for April 2020 and May 2020 forecasts by fitting a 2-way interaction MLM with time and domain as level 1 predictors and field of research as level 2 along with all 2 and 3-way interaction terms. We included the following covariates in the model: age, gender, and level of expertise. We then performed pairwise comparisons between psychology and other disciplines in *R* using *emmeans* package for each domain and each time point (Table S12 & Table S13). The results indicated no significant difference in the predictions made by psychologists versus those in other disciplines, after applying Benjamini-Hochberg correction to control for false discovery rate, *ps* > .10.

Comparing lay and academic retrospective estimates made in October/November of 2020

In October and November of 2020, we asked participants to look back and estimate how much they thought certain domains had changed in the last six months. To test whether a lack of differences between psychological scientists and lay people can be attributable to demographic differences, we fit a 2-way linear mixed model with sample (lay people vs. academics), domain, and their interaction as predictors of estimates, while controlling for ethnicity, political affiliation, age, gender, and income, and nesting observations in participants. Figure S10 present estimates from these models, showing close to identical results for models with and without covariates.

To test the difference between lay people's and psychological scientists' retrospective estimates in October/November 2020, we fit a 2-way linear mixed model with domain as level 1 predictor and sample dummy variable (lay people vs. psychological scientists) as level 2 predictor, along with the domain x sample interaction term. Then, we performed pairwise comparisons between lay people and psychological scientists in *R* using *emmeans* package for each domain, and subsequently used Benjamini-Hochberg method for false discovery rate correction to account for number of tests (Table S14). Lay people and psychological scientist only significantly disagreed in their estimates in five out of fifteen domains: generalized trust, delay of gratification, violence rates, individualism and depression.

In addition to this frequentist approach, we conducted a Bayesian analysis to test for statistical equivalence between psychological scientists' and lay people's retrospective estimates. The models were fit using *stan_glm* function from *rstanarm* package in *R*. For each domain, we fit a Bayesian linear mixed model with sample as the sole predictor and observations nested in participants, with normally distributed priors, $N \sim (0, 5)$, for predictor and intercept. Bayes Factor was computed in favor of the null hypothesis, such that there is no difference between lay and expert retrospective estimates (BF01; Table S15).

Finally, to test whether domain expertise affected psychological scientists' retrospective estimates we fit a 2-way linear mixed model with domain expertise (yes/no) × domain interaction predicting retrospective estimates. Observations were nested in participants. Fitted means and CIs were then extracted from the model and plotted (Figure S11). Having domain expertise had no significant impact upon retrospective estimates of change (see Table S16).

Effects of vividness of memories and news exposure on retrospective estimates of change

As a supplementary analysis, we examined how two characteristics: i) vividness of memories; ii) news exposure impacted retrospective assessments of scientists and lay people. We fit two linear mixed models to test whether vivid memories and news reports affected retrospective estimates for both psychological scientists and lay people by domain. In both models we included main effects of sample (psychological scientists vs. lay) and either vividness (vivid/not vivid) or news exposure (news/no news) and a sample × vividness / news exposure interaction term. Observations were nested in participants. We performed pairwise comparisons (news/no news or vivid/not vivid) with a subsequent Benjamini-Hochberg false discovery rate correction. As Table S17 shows, we observed systematic effects of vividness and news exposure resulting in more extreme estimates among both groups.

Extremity of prospective vs. retrospective estimates

To test whether retrospective estimates were more or less extreme than prospective estimates we fit two linear mixed models: one comparing retrospective against prospective estimates for psychological scientists and the other for lay people. In both models we included estimate type (prospective vs. retrospective), domain, and estimate type × domain interaction, while nesting observations in participants. Then, we performed domain-wise pairwise comparisons between prospective and retrospective estimates, applying Benjamini-Hochberg method for false discovery rate correction. Table S18 shows estimates and 95% CIs. Table S19 shows results of equivalent analyses with demographic covariates (ethnicity, political affiliation, age, gender, and income), suggesting that including covariates leads to largely identical results to those without covariates.

Forecasts and accuracy for psychologists vs. other scientists

We sought to determine whether psychologists were more or less accurate than other social/life scientists. As Figure S12 shows, for most estimates, groups were not significantly different from each other, zs < 2.12, ps < .172, with two exceptions: psychologists were somewhat less inaccurate than other social/life scientists for loneliness, z = 3.18, p = .011, and social/life scientists were less inaccurate than psychologists when predicting violent crimes, z = 3.51, p = .007. However, in both cases both groups were still inaccurate, and for loneliness most psychologists and other scientists predicted change in the opposite direction from the ground truth.

Study 4

Valuation of expert judgment by the general public

In Study 4, we examined valuation of expert judgments by the general public. Whereas scientists with expertise in psychology and public health were viewed as most likely to accurately estimate societal trends in depression, life satisfaction, loneliness, violence and related domains in the next half a year, politicians were viewed as least likely to estimate societal trends correctly (see Figure 4). Linear mixed model analyses showed a significant difference between groups, χ^2 (*df* =*9, N* = 203) = 1008.5, $p < .001$. Post-hoc pairwise comparisons with Tukey-correction showed that psychological scientists were viewed as significant more accurate than most groups, $5.52 < t\text{s} \leq 17.02$, $p\text{s} < .001$, with the exception of no difference from medicine, social work, and public health. Two additional observations stood out. First, though general public considered all groups of scientists and practitioners as more likely to be accurate in their predictions than an average American, $3.26 < \text{ts} \leq 16.94$, $p\text{s} < .038$, there was one group – politicians who were viewed as even less likely to accurately estimate societal change, $\underline{t} = 4.91$, $p < .001$. Second, experts in economics and political science were viewed as less likely to be accurate than all other groups of scientists and practioners, $3.66 < t\text{s} \leq 13.68$, $p\text{s} < .010$.

We also assessed the general public's preferences for groups they would be most likely to seek recommendations for the same societal issues. The results largely mirrored those for perception of prediction accuracy. Linear mixed model analyses showed a significant difference between groups, χ^2 (*df* =*9, N* = 203) = 1213.8, $p < .001$. Post-hoc pairwise comparisons with Tukey-correction showed that psychological scientists were viewed as significant more preferred to provide recommendations than most groups, $3.31 < t\text{s} \leq 23.93$, $p\text{s} < .031$, with the exception of no difference from social work and public health. Again, whereas all groups of scientists and practitioners were more preferred to provide recommendations than an average American, politicians were less preferred than an average American. Also, experts in economics and political science were viewed as less preferred to provide recommendations than all other groups of scientists and practioners, with the exception of historians.

We also asked participants to rank their top 3 preferred professions to provide an idea how the COVID-19 pandemic will affect human behavior and society in the long-term. Generalized linear mixed model analyses with a Poisson distribution showed a significant difference between groups, χ^2 (*df* =*9, N* = 203) = 533.1, $p < .001$. Post-hoc pairwise comparisons with Tukey-correction showed that psychological scientists were ranked as significantly more desirable than all groups, $7.16 < t\text{s} \leq 12.30$, $p\text{s} < .001$, with the exception of public health. Whereas most scientists and practitioners were ranked as more desirable than an average American, there was no difference in ranking position of an average American and a scientist with expertise in economics, whereas politicians and political scientists were ranked *lower* than an average American.

Valuation of expert judgment by academics / policy-makers

In addition to a sample of lay people from the general public, we also examined preferences for predictions and policy recommendations in a small sample of academics and policy-makers in Study 4b. As Figure S13 shows, the results were similar to Study 4a, with academics and policy makers favoring scientists with expertise in public health, social workers, and scientists with expertise in psychology the most, and politicians the least. When asked to choose the top three groups of experts to get a good idea about how the COVID-19 pandemic will affect human behavior and society in the long-term, scientists with expertise in psychology were similarly among the most frequently selected groups, with statistical analyses showing that psychological sicentists were selected significantly more often than other groups, $z = 3.75$, $p < .001$, $R^2 = .02$.

References

1. Trope Y & Liberman N (2010) Construal-level theory of psychological distance. *Psychol Rev* 117(2):440-463.
2. Sherman SJ, Cialdini RB, Schwartzman DF, & Reynolds KD (1985) Imagining can heighten or lower the perceived likelihood of contracting a disease: The mediating effect of ease of imagery. *Pers Soc Psychol Bull* 11(1):118-127.
3. Jaidka K*, et al.* (2020) Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci. USA* 117(19):10165-10171.
4. Schwartz H*, et al.* (2013) Characterizing geographic variation in well-being using tweets. *Proc. Intl. AAAI Conf. on Web and Social Media* 7:583-591.
5. Tausanovitch C, Vavreck L, Reny T, Hayes AR, & Rudkin A (2019) Democracy fund+ UCLA nationscape methodology and representativeness assessment. *Democracy Fund Voter Study Group.* https://www*. voterstudygroup. org/uploads/reports/Data/NS-Methodology-Representativeness-Assessment. pdf*.

6. Mercer AW (2018) Selection bias in nonprobability surveys: A causal inference approach. (University of Maryland, College Park).
7. Boxell L, Conway J, Druckman JN, & Gentzkow M (2020) Affective polarization did not increase during the coronavirus pandemic. (National Bureau of Economic Research).
8. Kim HS, Sherman DK, & Updegraff JA (2016) Fear of Ebola: The influence of collectivism on xenophobic threat responses. *Psychol Sci* 27(7):935-944.
9. Aarøe L, Osmundsen M, & Petersen MB (2016) Distrust as a disease avoidance strategy: Individual differences in disgust sensitivity regulate generalized social trust. *Front. Psychol.* 7:1038.
10. Dunwoody PT & Funke F (2016) The Aggression-Submission-Conventionalism Scale: Testing a new three factor measure of authoritarianism. *J Soc Polit Psychol* 4(2):571-600.
11. Rosenfeld R & Lopez E (2020) Pandemic, social unrest, and crime in US Cities. *Council on Criminal Justice*.
12. Leitner JB, Hehman E, Ayduk O, & Mendoza-Denton R (2016) Blacks' death rate due to circulatory diseases is positively related to whites' explicit racial bias: A nationwide investigation using project implicit. *Psychol Sci* 27(10):1299-1311.
13. Ofosu EK, Chambers MK, Chen JM, & Hehman E (2019) Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proc Natl Acad Sci USA* 116(18):8846-8851.

14. Hehman E, Calanchini J, Flake JK, & Leitner JB (2019) Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *J Exp Psychol: Gen* 148(6):1022-1040.

15. Greenwald AG, Nosek BA, & Banaji MR (2003) Understanding and using the implicit association test: I. An improved scoring algorithm. *J Pers Soc Psychol* 85(2):197-216.

Supplemental Figures

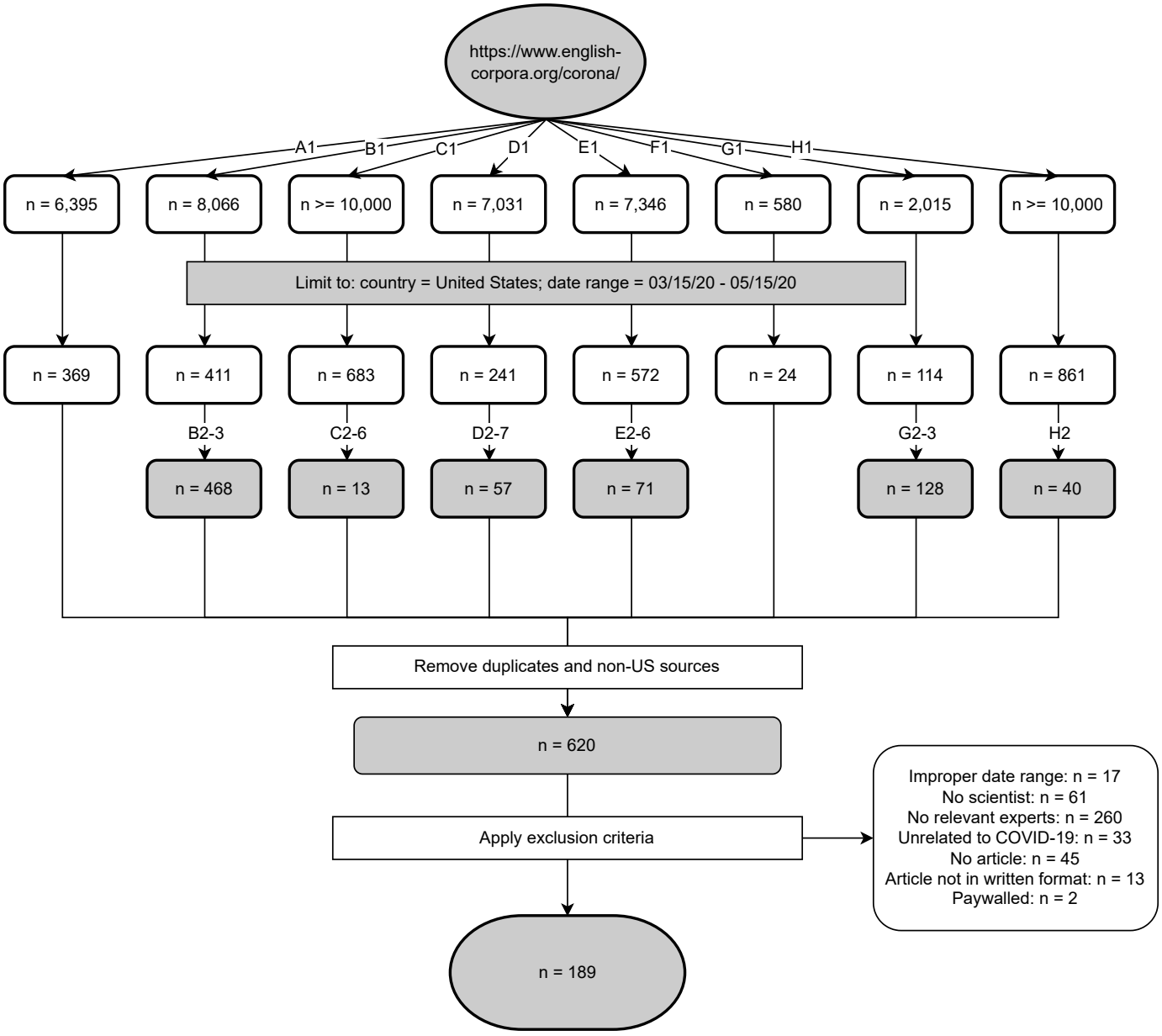


Figure S1. Identification and selection of media interview texts.

Texts were derived from The Coronavirus Corpus. This corpus operates by creating "mini-corpora" that contain all texts in which a single target word appears. Corpora containing conjunctions of words (e.g., "psychology professor") can be created by searching within mini-corpora for texts with an additional word. To identify texts, we thus first created mini-corpora of texts that were published between March 15 and May 15, 2020, and included a single keyword (A1 = "psychologist", B1 = "psychology", C1 = "psychological", D1 = "cognitive", E1 = "behavioral", F1 = "neuroscientist", G1 = "neuroscience", H1 = "scientist"). For domains B, C, D, E, G, and H, we narrowed the list further by including a second limiting term: e.g., B1 = "psychology" + B2 = "professor", B1 = "psychology" + B3 = "researcher", C2 = "sciences", C3 = "sciences", C4 = "scientist", C5 = "researcher", C6 = "professor of", D2 = "science" , D3 = "sciences" , D4 = "scientists" , D5 = "scientist", D6 = "professor of" , D7 = "researcher" , E2 = "science", E3

= “sciences”, E4 = “scientist”, E5 = “professor of”, E6 = “researcher”, G2 = “professor of” , G3 = “researcher”. Duplicates and texts from non-US sources were then removed, and exclusion criteria were applied, leaving a total of 189 texts, some with interviews with multiple experts on multiple topics.

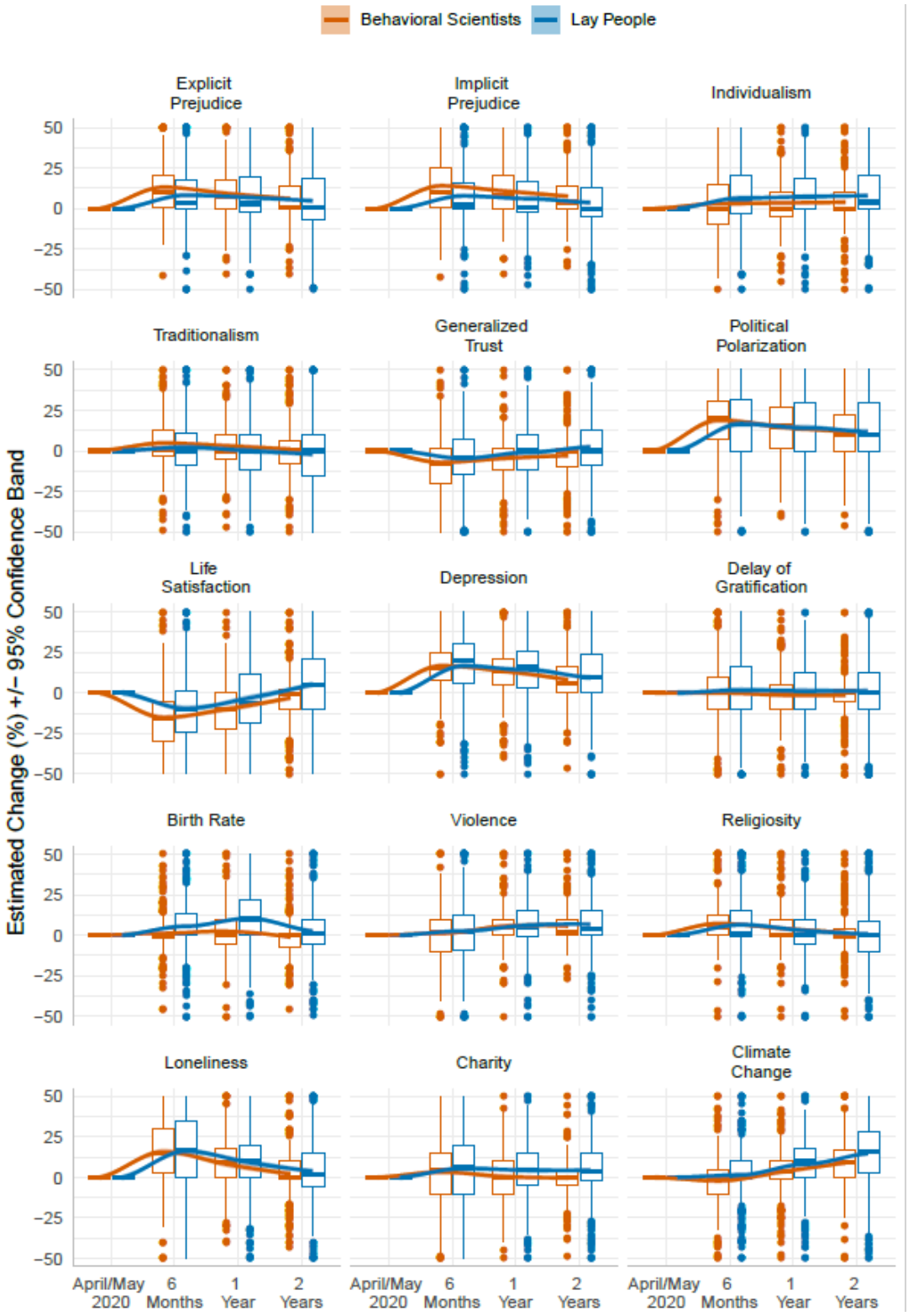


Figure S2. Predictions for change across 15 societal domains from psychological scientists and lay people. Graphs indicate boxplots for a given time-point forecast (half a year, year, two years from April/May 2020) and loess line of best fit across time points with 95% confidence bands around the estimate.

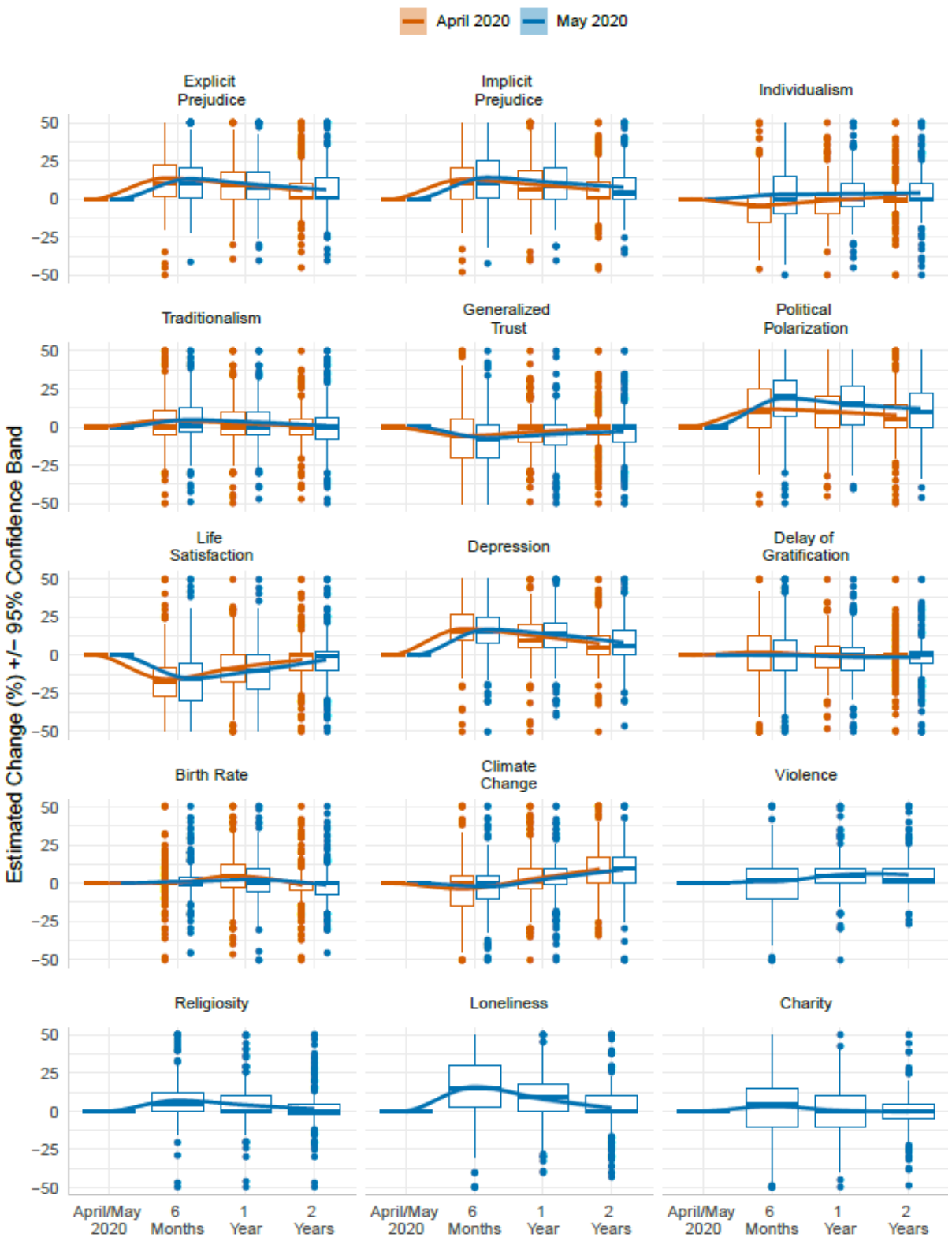


Figure S3. Psychological scientists predicting societal change in April vs. May, 2020. Graphs indicate boxplots for a given time-point forecast (half a year, year, two years from April/May 2020) and loess line of best fit across time points with 95% confidence bands around the estimate. Positive numbers refer to positive change, whereas negative—a negative change.

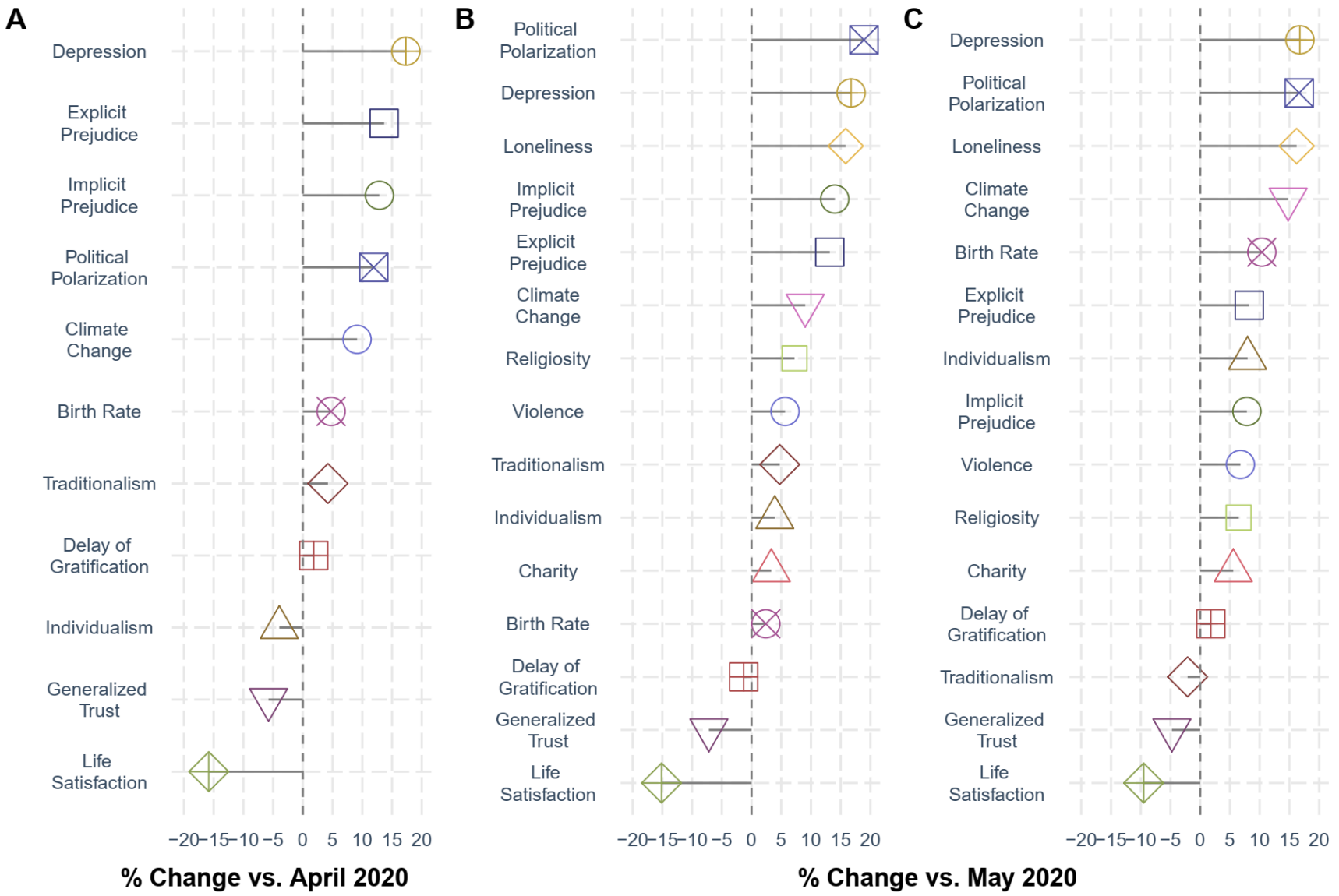


Figure S4. Ranking of domains based on magnitude and direction of predicted societal change, as estimated by psychological scientists over two years from (A) April 2020 (B) and May 2020 as well as (C) lay people's predictions over the same time period.

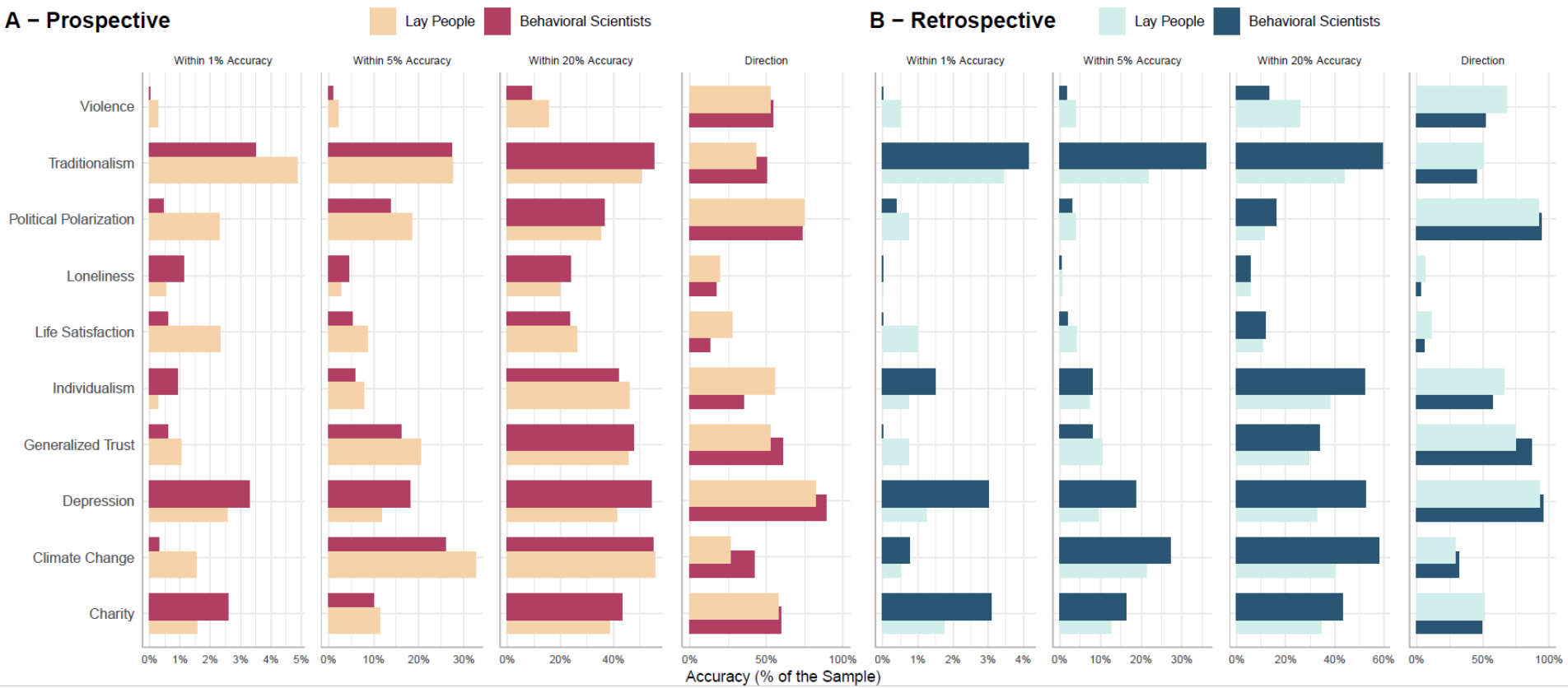


Figure S5. Percentage of a given sample that accurately estimated societal change. Panels from left to right represent different accuracy benchmarks: percentage of accurate estimates that (*i*) fall within 1%, (*ii*) 5% and (*iii*) 20% of the accuracy and (*iv*) directional accuracy of the trend.

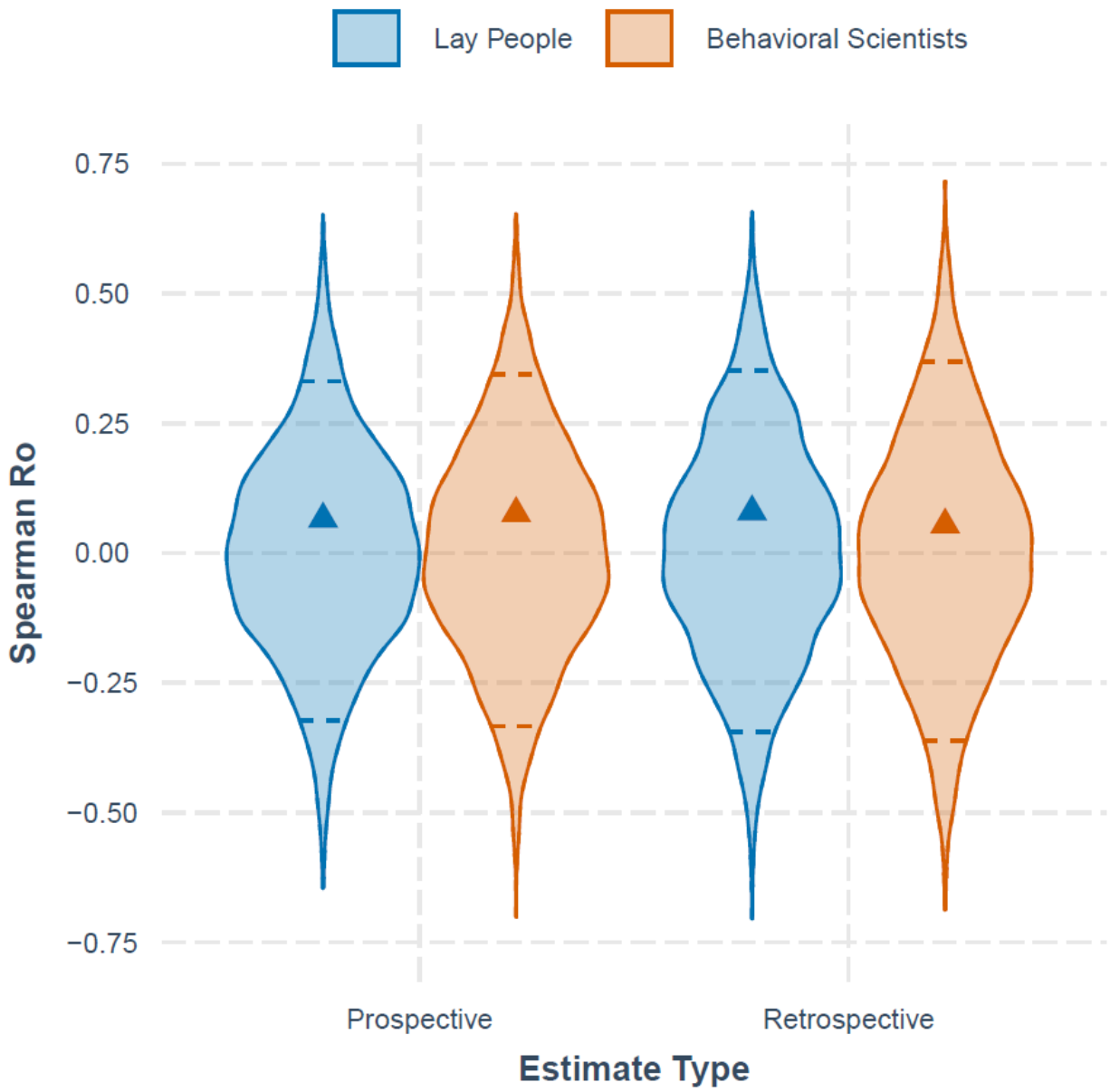


Figure S6. Rank-order accuracy of estimates. Triangles represent the average Spearman's rank order correlation between individual estimates and observed change across domains. Violin plots represent the expected null distribution of group-averaged correlation coefficients, constructed using 5000 permutations that randomly shuffled domain labels of the observed outcomes. Dashed lines indicate the 95% confidence interval of the null distribution. Average rank-order estimates of each group fell well within this distribution, suggesting that they are not significantly different from estimates expected by chance.

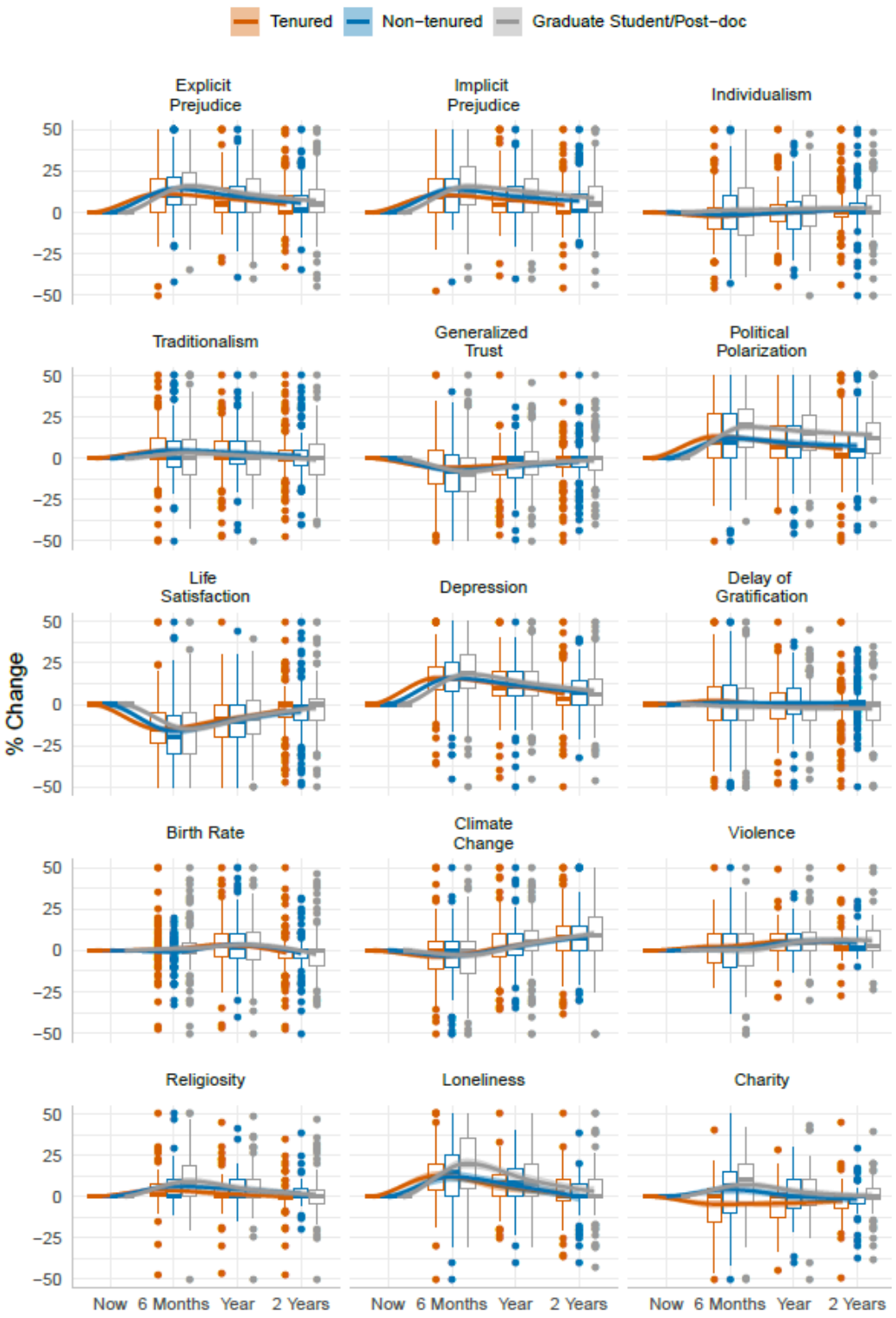


Figure S7. Psychological scientists' predictions for change across 15 societal domains by Faculty Type. Data is pooled across Studies 3a-b. Graphs indicate boxplots for a given time-point forecast (half a year, year, two years from April/May 2020) and loess line of best fit across time points with 95% confidence bands around the estimate. Positive numbers refer to positive change, whereas negative—a negative change.

Figure S8. Psychological scientists' predictions for change across 15 societal domains by area of expertise from April/May 2020. Graphs indicate means and 95% confidence intervals for a given time-point forecast (half a year, year, two years from April/May 2020). Positive numbers refer to positive change, whereas negative refer to a negative change

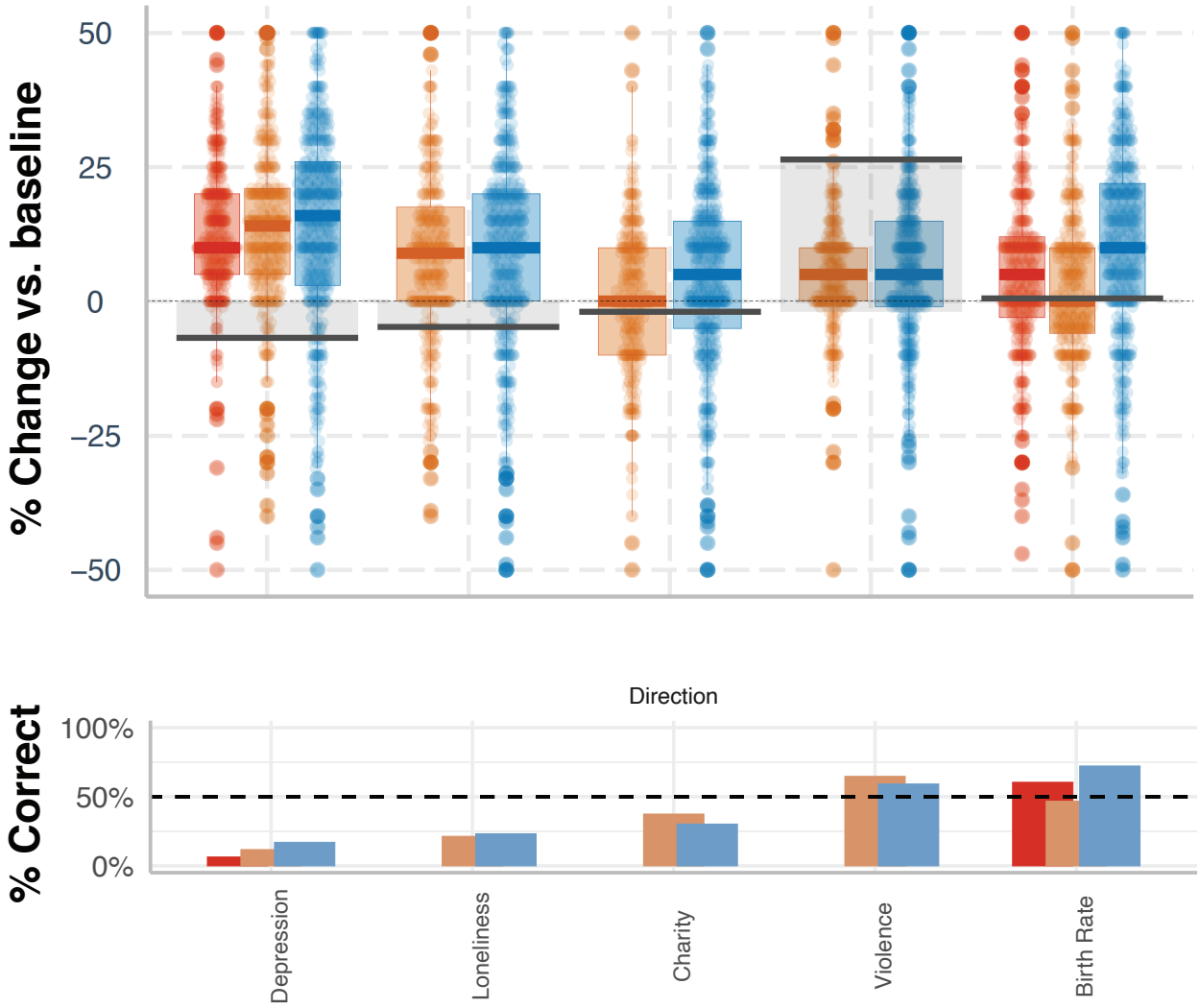


Figure S9. The accuracy of 12-month prospective judgments (made in April and May 2020) of societal change. Predictions, along with objective markers for five available domains, are displayed for prospective judgments in psychological scientists and laypeople. Box-plots show median and 25/75% confidence intervals. Accuracy (measured as directionally correct predictions) is displayed below. Prospective data includes two separate samples of psychological scientists surveyed in late March/early April and late April/early May).

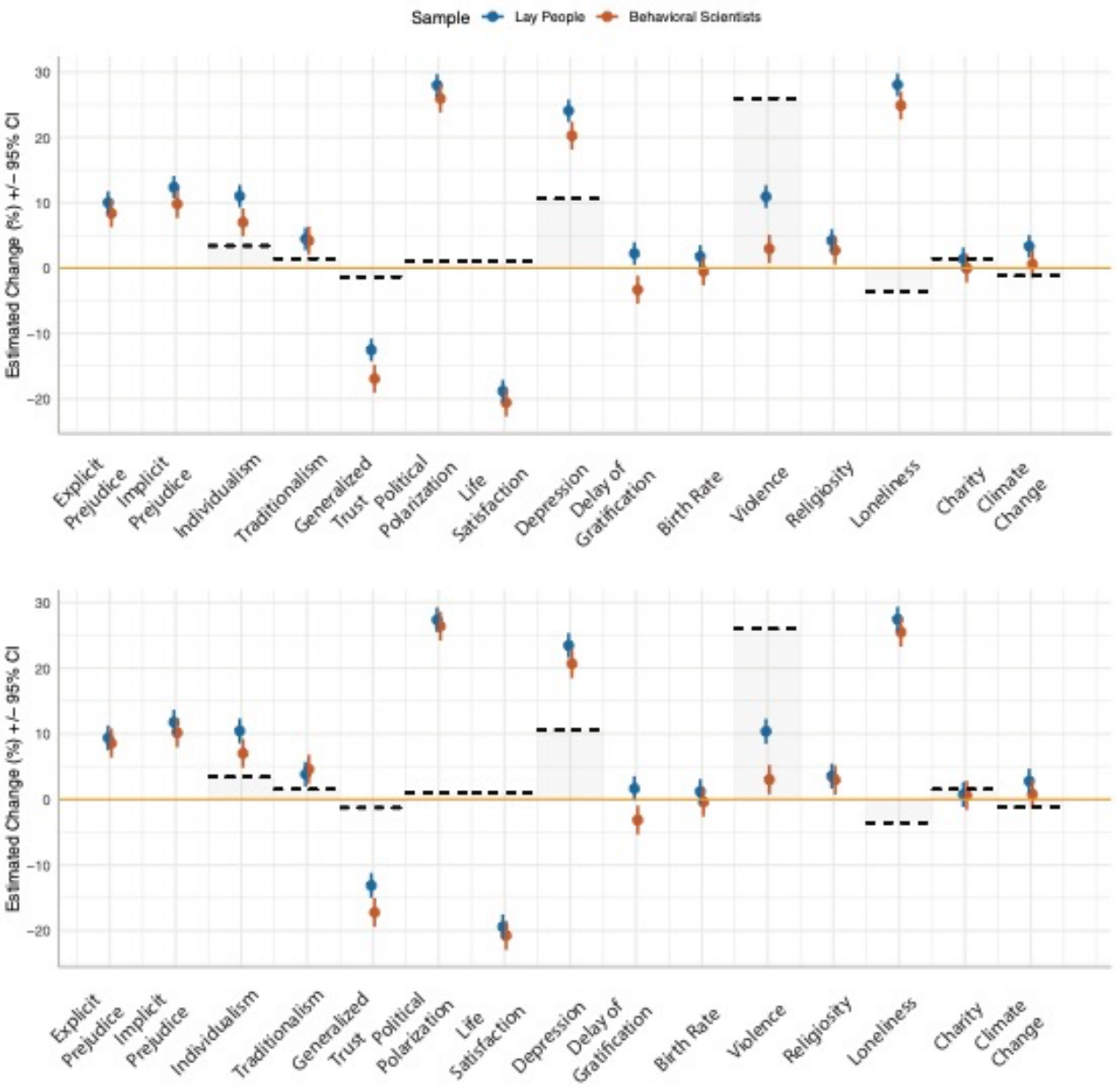


Figure S10. Retrospective mean estimates of social change (i.e., estimated change from April to October/November 2020) by lay people and psychological scientists. Gray bars display ground truth estimates for available domains. Top: without covariates; Bottom: controlling for ethnicity, political affiliation, age, gender, and income. Error bars show 95% CI

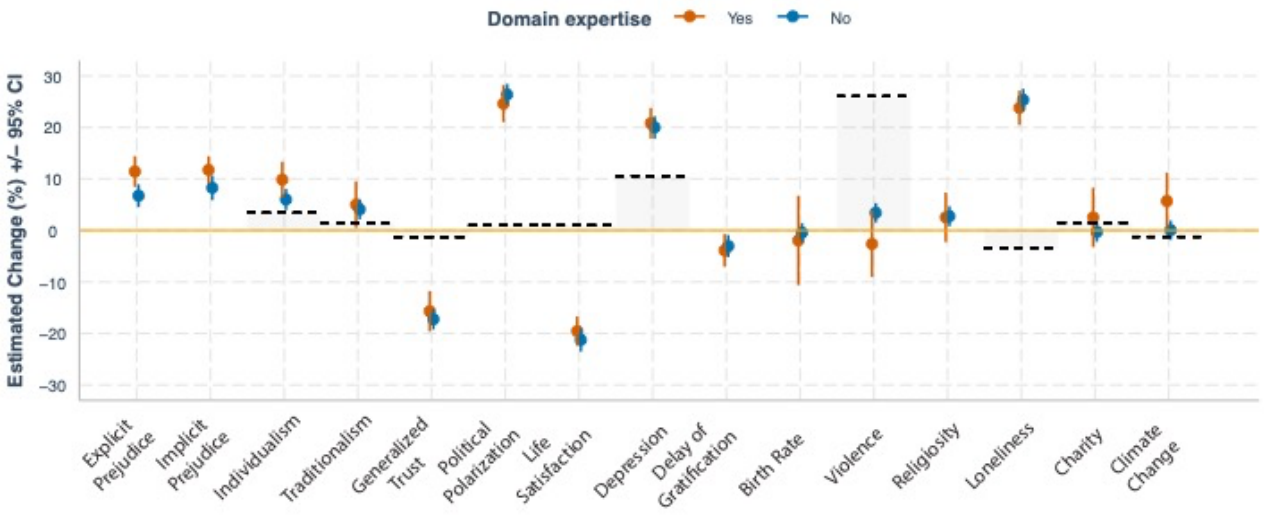


Figure S11. Retrospective mean estimates of change across 15 societal domains for psychological scientists by domain expertise. Gray bars display ground truth estimates for available domains.

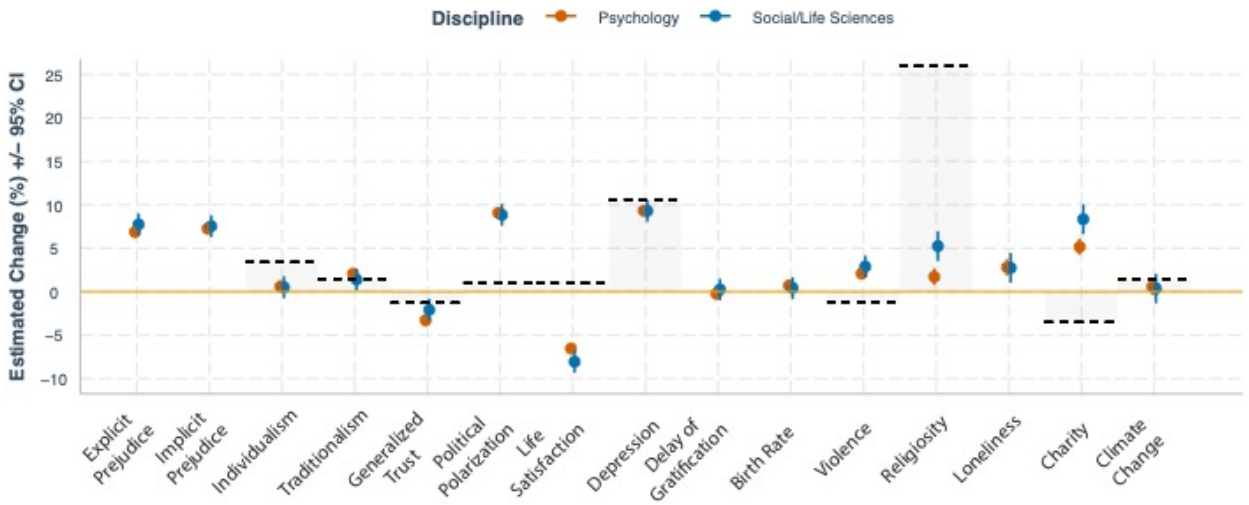


Figure S12. Forecasting estimates (Study 3a) and accuracy benchmarks for behavioral scientists from psychology vs. other disciplines. Gray bars display ground truth estimates for available domains.

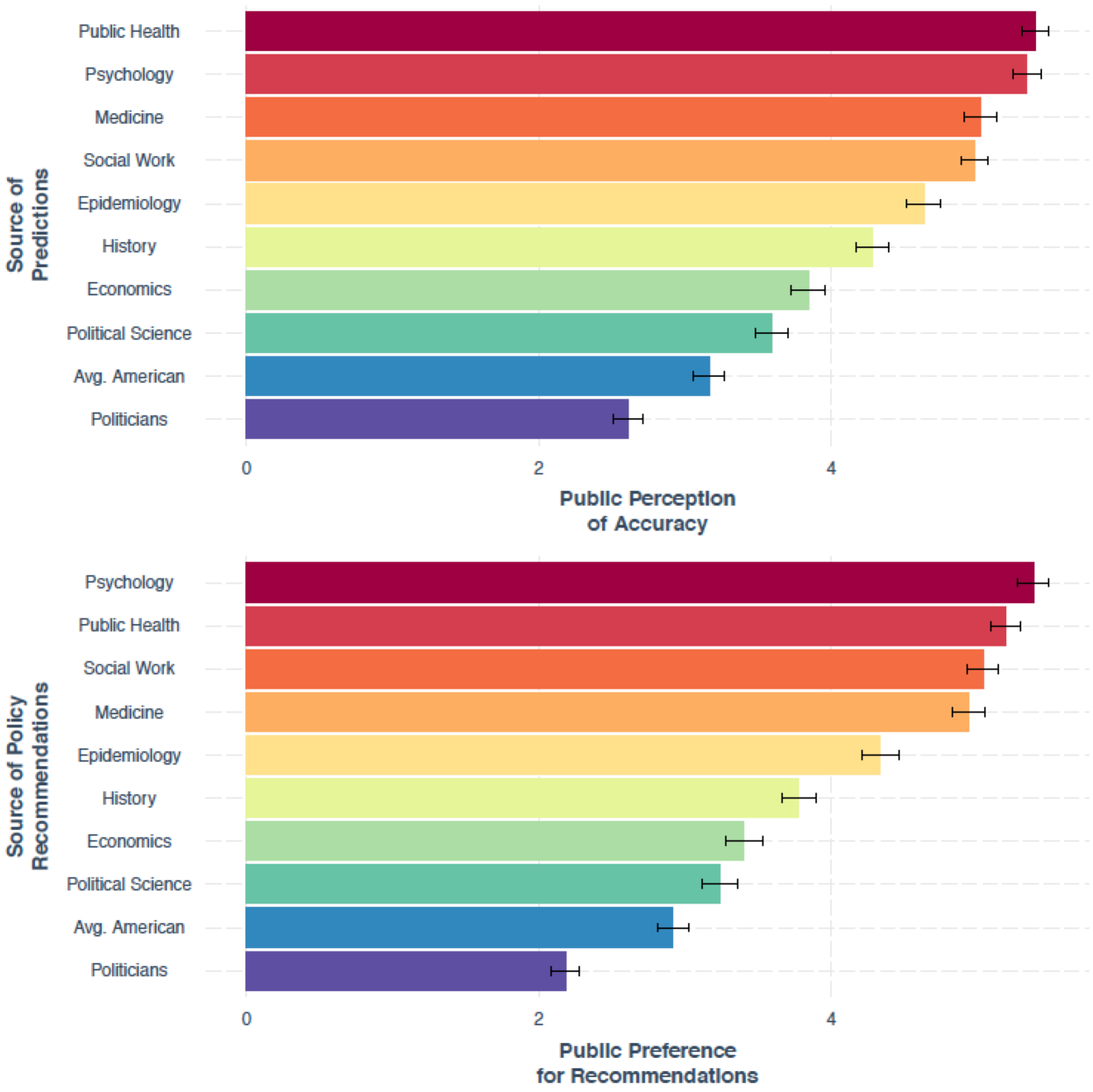


Figure S13. Perception of experts' accuracy when predicting societal change over COVID-19 (top) and preferences for experts who academics and policy-makers (as opposed to lay individuals) favor to provide recommendations concerning societal issues concerning depression, life satisfaction, loneliness, violence and related issues resulting from the pandemic (bottom).

Supplemental Tables

Table S1. Descriptive Statistics of the Studies 3-4 (Forecasting and Retrospective Assessment of Societal Change)

		Study 3a	Study 3b (Psychological scientists)	Study 3c (Lay people)	Study 4a (Psychological scientists)	Study 4b (Lay people)	Society for Personality and Social Psychology
Sample (*N*)		401	316	394	270	411	
Age	M_{age}	41	39	45	38	45	-
	Range	22 – 88	19 – 87	18-78	22-76	18-78	
Gender	% Female	45%	63%	52%	72%	50%	54%
Household Income (*Md*)			$100,001 - $150,000	$50,001 - $75,000	$75,001 - $100,000	$50,001 - $75,000	-
Ethnicity	White		234 (75%)	269 (68%)	212 (79%)	276 (67%)	62%
	Asian		31 (10%)	29 (7%)	22 (8%)	33 (8%)	14%
	Hispanic		13 (4%)	18 (5%)	5 (2%)	21 (5%)	5%
	Black		2 (1%)	51 (13%)	3 (1%)	57 (14%)	4%
	Middle Eastern		3 (1%)	2 (1%)	3 (1%)	4 (1%)	2%
	East Indian		8 (3%)	-	3 (1%)	-	-
	Aboriginal		-	2 (1%)	3 (1%)	-	-
	Other		21 (7%)	23 (6%)	18 (7%)	18 (4%)	1%
Country	USA	195 (50%)	194 (62%)	394 (100%)	196 (73%)	411 (100%)	77%
	Canada	93 (24%)	57 (18%)		47 (17%)		7%
	Germany	17 (4%)	7 (2%)		5 (2%)		2%
	United Kingdom	16 (4%)	17 (5%)		4 (1%)		2%
	Australia	9 (2%)	3 (1%)		-		1%
	Netherlands	7 (2%)	3 (1%)		1 (< 1%)		1%
	Switzerland	6 (2%)	3 (1%)		4 (1%)		1%
	China	2 (< 1%)	5 (2%)		-		-
	Other	48 (12%)	23 (8%)		13 (5%)		9%
Research Field	Psychology	330 (83%)	237 (76%)				
	Neuroscience	8 (2%)	10 (3%)				
	Political Science	-	21 (7%)				
	Economics	11 (3%)	8 (3%)				
	Computer Science	1 (< 1%)	5 (2%)				
	Sociology	3 (1%)	3 (1%)				
	Biology	-	3 (1%)				
	Medicine and Epidemiology	2 (<1%)	3 (1%)				
	Other	44 (11%)	26 (8%)				
Academic Position	Tenured Faculty	137 (34%)	82 (26%)				36%
	Non-Tenured Faculty	73 (19%)	37 (12%)				9%
	Postdoctoral Researchers	40 (10%)	32 (10%)				n/a
	Graduate Students	93 (23%)	102 (33%)				37%
	Research Scientists	28 (7%)	29 (9%)				n/a
	Other	27 (7%)	34 (11%)				15%
Organization Type	College/University	368 (92%)	280 (89%)	64 (16%)	259 (96%)	56 (14%)	
	Government	10 (3%)	10 (3%)	23 (6%)	2 (1%)	21 (5%)	
	Private Company	15 (4%)	15 (5%)	164 (42%)	6 (2%)	174 (43%)	
	Self-employed	7 (2%)	8 (3%)	140 (36%)	2 (1%)	154 (38%)	
Organization Size	< 10	6 (2%)	10 (3%)	156 (40%)	3 (1%)	154 (39%)	

	11 - 100	11 (3%)	7 (2%)	54 (14%)	5 (2%)	49 (12%)
	101 - 1,000	17 (4%)	19 (6%)	55 (14%)	7 (3%)	79 (20%)
	1,001 – 10,000	86 (22%)	52 (17%)	53 (14%)	43 (16%)	57 (14%)
	10,001 – 50,000	201 (51%)	163 (52%)	39 (10%)	139 (52%)	43 (11%)
	50,000+	74 (19%)	61 (20%)	33 (8%)	72 (27%)	16 (4%

Notes. Due to a technical error, information about the country of residence and research field was not collected in Study 4. To estimate country of residence for Study 4, we used each person's IP address to geolocate their country of residence at the time of taking the survey (using the *rgeolocate* package in *R*). Statistics for comparison to demographics of the Society for Personality and Social Psychology come from https://spsp.org/sites/default/files/Member-Diversity-Statistics-December-2019.pdf.

Table S2. Variable descriptions provided to participants for the fifteen domains.

Variable	Description Provided to Participant
Generalized Trust	For each time period below, use the sliders to indicate how much generalized trust in other people in the United States will change (in %) from where it stands right now.
Life Satisfaction	Consider a person's life satisfaction in the United States, an overall assessment of how content a person is with their life overall, and measured by endorsement of statements like "The conditions of my life are excellent."
Clinical Depression	Consider clinical depression in the United States, as diagnosed by criteria listed in the Diagnostic and Statistical Manual, is characterized by feeling sad, losing interest in activities once enjoyed, and a loss of energy over a prolonged period of time, and is measured by agreement with statements like "I am sad all the time and I can't snap out of it."
Political Affective Polarization	Consider political affective polarization in the United States, defined as the degree of dislike and distrust towards those from the opposing political party.
Individualism	Consider people's concern for individualism in the United States, defined as values that emphasize the uniqueness, autonomy and individual goal pursuit.
Traditionalism	Consider people's concern for traditionalism in the United States, defined as a concern for adherence to traditional beliefs and practices, and measured by endorsement of items like, "the 'old-fashioned ways' and 'old-fashioned values' still show the best way to live."
Delay of Gratification	Consider the extent to which people delay their gratification in the United States, defined as resistance to the temptation of an immediate reward in preference for a later, larger reward.
Explicit Prejudice	Consider endorsement of general levels of explicit prejudice toward ethnic minorities in the United States, defined in this case as consciously holding negative attitudes toward ethnic or racial minorities, and measured by endorsement of items such as "Over the past few years, ethnic/racial minorities have gotten more economically than they deserve."
Implicit Prejudice	Consider measures of implicit prejudice toward ethnic minorities in the United States, defined here as negative feelings and/or beliefs about an ethnic group that people hold without being aware of it, and which is thought to operate automatically, with little intention or control on the part of the person.
Concern for Climate Change	Consider Americans' concern about climate change in the United States, as measured by endorsement of the idea that global warming is personally worrisome and a problem that is of pressing concern.
Birth Rates	Consider birth rates in the United States, defined as the total number of live births per 1000 women in the total population.
Charitable Giving	Consider charitable giving in the United States, defined as donations made by individuals to non-profit organizations, charities, or private foundations.
Violent Crimes	Consider violent crimes in the United States, defined as any violation of the law that is committed with physical force and is measured by rates of murder, manslaughter, rape, robbery, and aggravated assault.
Religiosity	Consider religiosity in the United States, defined as belief in a higher power, and measured by questions like "How certain are you of the existence of a higher power?"
Loneliness	Consider loneliness in the United States, defined as a person's subjective feeling of loneliness, and measured by agreement with statements like "A lot of times I feel lonely." and "I often feel left out of things."

Other Variable	Is there another key psychological or social issue in the United States we have not mentioned that you think would change? If so, please identify ONE key issue that you think is most important.

Table S3. Sources for benchmarking accuracy for estimates of societal change.

Dimensions	Source	Reference$_{Period}$	Change$_{Period}$	N$_{Reference}$	N$_{Change}$	Observed Change
Life	Gallup Panel – COVID-19 Survey	Apr 23 - May 5 (2020)	Oct 14 - Oct 26 (2020)	10,058	2,667	1%
Satisfaction	ICL/Yougov – Covid 19 Behaviour Tracker	Apr 27 – May 3 (2020)	Sep 14 – Sep 20 (2020)	1,003	966	-3.2%
	Twitter-based estimates	Apr 01 - Apr 30 (2020)	Oct 01 - Oct 31 (2020)	-	-	-0.3%
Loneliness	Understanding America Study	Apr 15 – May 15 (2020)	Sep 15 – Oct 12 (2020)	5,321	5,279	-3.5%
	Gallup Panel– COVID - 19 Survey	Apr 23 - May 5 (2020)	Oct 14 - Oct 26 (2020)	10,071	2,659	-16%
Depression	Household Pulse Survey	Apr 23 - May 5 (2020)	Oct 14 - Oct 26 (2020)	69,316	76,034	10.6%
	Understanding America Study	Apr 15 – May 15 (2020)	Sep 15 – Oct 12 (2020)	5,614	5,277	-1.9%
	Gallup Panel - COVID-19 Survey	May 11- May 30 (2020)	Oct 14 – Oct 26 (2020)	11,175	2,666	3%
Affective Polarization	Nationscape	Apr 23 - May 6 (2020)	Oct 01 - Oct 28 (2020)	8,108	26,000	1%
	Gallup U.S. Poll**	Apr 01 - Apr 30 (2020)	Oct 01 - Oct 31 (2020)	-	-	6.4%
Individualism	COVID-19 Attitudes Survey	Apr 22 – Apr 24 (2020)	Sep 23 – 28 (2020)	1,510	805	3.4%
Generalized Trust	COVID-19 Attitudes Survey	Apr 22 – Apr 24 (2020)	Sep 23 – 28 (2020)	1,510	805	-1.3%
Traditionalism	Nationscape	Apr 23 - May 6 (2020)	Oct 01 - Oct 28 (2020)	13,058	26,222	1.5%
	COVID-19 Attitudes Survey	Apr 22 – Apr 24 (2020)	Sep 23 – 28 (2020)	1,510	805	5%
Violence	Pandemic, Social Unrest, and Crime in U.S. Cities (November 2020)	Apr 01 - Apr 30 (2020)	Oct 01 - Oct 31 (2020)	-	-	26%
Climate Change	Nationscape	Apr 23 - May 6 (2020)	Oct 01 - Oct 28 (2020)	13,069	26,376	-1.2%
Charitable Giving	Giving Tuesday*	April-May (2020)	Oct – Nov (2020)	17.39%	17.66%	1.5%
Explicit Prejudice	*Project Implicit*	*Apr 01 – Apr 30 (2020)*	*Oct 01 – Oct 31 (2020)*	*70,700*	*239,740*	*-123%*
Implicit Prejudice	*Project Implicit*	*Apr 01 – Apr 30 (2020)*	*Oct 01 – Oct 31 (2020)*	*70,700*	*239,740*	*- 6.7%*
		12-Month Benchmarks				
Depression	Household Pulse Survey	Apr 23 - May 5 (2020)	Apr 28 – May 10 (2021)	69,316	65,513	-6.8%
Loneliness	Understanding America Study	Apr 15 – May 15 (2020)	Apr 15 – May 15 (2021)	5,321	4887	-4.8%
Violence	National Commission on COVID-19 and Criminal Justice	April 2020	April 2021	-	-	26.53%
Charitable Giving	Giving Tuesday	April 2020	April 2021	-	-	-1.9%
Birth Rate	Human Fertility Database	April 2020	April 2021	290,252	292,000	0.6%
		Alternative Benchmarks for Study 2a (early April time-period)				
Affective Polarization	Nationscape	Mar. 26-Apr. 8	Oct. 1-Oct. 7	12,123	5,890	1.14%
Life Satisfaction	Gallup	Apr. 1-15	Oct. 15-26	18,256	2,681	1.8%
Traditionalism	Nationscape	Mar. 26-Apr. 8	Oct. 1-Oct. 7	12,123	5,890	-1.5%†
Climate Change	Nationscape	Mar. 26-Apr. 8	Oct. 1-Oct. 7	12,123	5,890	-6.3%
Explicit Prejudice	*Project Implicit*	*Mar. 1-Mar. 30*	*Sept. 1.-Sep. 30*	*71015*	*220,073*	*-629%*
Implicit Prejudice	*Project Implicit*	*Mar. 1-Mar. 30*	*Sept. 1.-Sep. 30*	*71015*	*220,073*	*-14.5%*
Birth Rate (1-year)	Human Fertility Database	Mar 2020	Mar 2021	301,625	302,000	.1%

Note. *Giving Tuesday estimates indicate % of yearly total, thereby adjusting for yearly base rates. ** Gallup Poll data is based on nationally representative discrete multi-day surveys (https://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx). † Difference in sign from primary measure (late April – late Oct./early Nov. time period).

Table S4. Significance testing of the difference between forecasted estimates for April 2022 and baseline in April 2020 and for May 2022 and baseline May 2020.

Domain	*t*	*p*	*p* *
April 2022 vs. April 2020			
Explicit Prejudice	7.96	<.001	<.001
Implicit Prejudice	8.5	<.001	<.001
Individualism	2.33	0.02	0.028
Traditionalism	0.81	0.417	0.459
Generalized Trust	-1.89	0.059	0.072
Political Polarization	8.98	<.001	<.001
Life Satisfaction	-4.51	<.001	<.001
Depression	12.11	<.001	<.001
Delay of Gratification	-0.58	0.566	0.567
Birth Rate	-2.37	0.018	0.028
Climate Change	11.87	<.001	<.001
May 2022 vs. May 2020			
Explicit Prejudice	14.78	<.001	<.001
Implicit Prejudice	14.63	<.001	<.001
Individualism	4.85	<.001	<.001
Traditionalism	5.11	<.001	<.001
Generalized Trust	-7.25	<.001	<.001
Political Polarization	16.86	<.001	<.001
Life Satisfaction	-13.69	<.001	<.001
Depression	16.41	<.001	<.001
Delay of Gratification	-1.67	0.097	0.1
Birth Rate	2.76	0.006	0.007
Climate Change	9.82	<.001	<.001
Violence	7.41	<.001	<.001
Religiosity	8.34	<.001	<.001

Loneliness	12.19	<.001	<.001
Charity	2.82	0.005	0.006

Note: p-values in the rightmost column were adjusted using the Benjamini-Hochberg false discovery rate correction.

Table S5. Significance Testing Whether Forecasted Change for Each Domain (Linear and Quadratic Temporal Trends) Vary by Group of Sampled Psychological scientists (April vs. May 2020).

Domain	Effect	*t*	*p*	*p**
Explicit Prejudice	Linear	1.11	.266	.568
	Quadratic	0.50	.618	.849
Implicit Prejudice	Linear	0.64	.525	.825
	Quadratic	0.17	.869	.893
Individualism	Linear	-4.25	<.001	.001
	Quadratic	1.02	.310	.568
Traditionalism	Linear	-0.19	.854	.893
	Quadratic	0.14	.893	.893
Generalized Trust	Linear	-0.44	.661	.856
	Quadratic	-0.15	.879	.893
Political Polarization	Linear	-2.75	.006	.045
	Quadratic	1.68	.093	.342
Life Satisfaction	Linear	-1.13	.258	.568
	Quadratic	2.11	.035	.157
Depression	Linear	1.46	.144	.412
	Quadratic	-1.02	.309	.568
Delay of Gratification	Linear	0.76	.448	.758
	Quadratic	-0.50	.616	.849
Birth Rate	Linear	-2.10	.036	.157
	Quadratic	3.36	.001	.009
Climate Change	Linear	-1.44	.150	.412
	Quadratic	0.17	.863	.893

Note: Rightmost *p* value column was adjusted for number of tests using Benjamini-Hochberg false discovery rate correction. For 3 temporal estimates, estimation of change can be parsimoniously decomposed into overall degree of change (i.e., linear effect) and possible curve in the estimated trajectory (i.e., quadratic effect). Therefore, in our analyses we focused on estimation of sample-wise differences in linear and quadratic effects.

Table S6. Significance Testing Whether Forecasted Change for Each Domain (Linear and Quadratic Temporal Trends) Vary between Lay People vs. Psychological scientists.

Dimension	Effect	*t*	*p*	*p**
Explicit Prejudice	Linear	-3.44	.001	.018
	Quadratic	1.40	.163	.543
Implicit Prejudice	Linear	-2.40	.017	.166
	Quadratic	.91	.363	.796
Individualism	Linear	-.43	.667	.854
	Quadratic	.15	.883	.883
Traditionalism	Linear	.43	.668	.854
	Quadratic	-.26	.799	.854
Generalized Trust	Linear	-2.29	.022	.168
	Quadratic	.22	.826	.854
Political Polarization	Linear	-1.89	.059	.300
	Quadratic	1.16	.248	.707
Life Satisfaction	Linear	-1.80	.072	.300
	Quadratic	.25	.799	.854
Depression	Linear	-1.13	.259	.707
	Quadratic	.37	.712	.854
Delay of Gratification	Linear	-.70	.486	.854
	Quadratic	.44	.663	.854
Birth Rate	Linear	-.86	.389	.796
	Quadratic	3.18	.002	.023
Climate Change	Linear	-1.75	.080	.300
	Quadratic	.26	.795	.854
Religiosity	Linear	.39	.695	.854
	Quadratic	-.55	.579	.854

Charity	Linear	-1.77	.076	.300
	Quadratic	.80	.425	.796
Violence	Linear	.55	.579	.854
	Quadratic	-.87	.383	.796
Loneliness	Linear	-.84	.400	.796
	Quadratic	.26	.798	.854

Note: Rightmost *p* values were adjusted for number of tests using Benjamini-Hochberg false discovery rate correction. For 3 temporal estimates, estimation of change can be parsimoniously decomposed into overall degree of change (i.e., linear effect) and possible curve in the estimated trajectory (i.e., quadratic effect). Therefore, in our analyses we focused on estimation of sample-wise differences in linear and quadratic effects.

Table S7. Comparisons of Prospective and Retrospective Estimates against Actual Change between April 2020 and October 2020 by Sample and Dimension.

Sample	Dimension	Estimate Type	Estimate	Actual Change	$M_{difference}$	*t*	*p*	*p**
Lay People	Depression	Prospective	16.76	10.6	6.16	6.00	<.001	< .001
		Retrospective	24.09	10.6	13.49	15.61	<.001	< .001
	Life Satisfaction	Prospective	-9.51	1	10.51	-9.39	<.001	< .001
		Retrospective	-18.83	1	19.83	-19.45	<.001	< .001
	Implicit Prejudice	Prospective	7.85	-6.7	14.55	16.46	<.001	< .001
	Explicit Prejudice	Retrospective	12.37	-6.7	19.07	19.78	<.001	< .001
		Prospective	8.25	-123	131.25	146.37	<.001	< .001
		Retrospective	10.04	-123	133.04	130.20	<.001	< .001
	Loneliness	Prospective	16.21	-3.5	19.71	16.45	<.001	< .001
		Retrospective	28.04	-3.5	31.54	33.95	<.001	< .001
	Political	Prospective	16.65	1	15.65	14.17	<.001	< .001
	Polarization	Retrospective	28.00	1	27.00	27.06	<.001	< .001
	Climate Change	Prospective	1.52	-1.2	2.72	3.33	.001	.001
		Retrospective	3.32	-1.2	4.52	4.63	<.001	< .001
	Traditionalism	Prospective	2.11	1.5	0.61	0.66	.512	.534
		Retrospective	4.48	1.5	2.98	3.02	.003	.004
	Violence	Prospective	2.61	26	23.40	-23.88	<.001	< .001
		Retrospective	10.93	26	15.07	-15.01	<.001	< .001
	Charity	Prospective	5.56	1.5	4.06	3.51	.001	.001
		Retrospective	1.37	1.5	0.13	-0.13	.899	.899
	Generalized Trust	Prospective	-4.76	-1.3	3.46	-3.53	<.001	< .001
		Retrospective	-12.54	-1.3	11.24	-11.03	<.001	< .001
	Individualism	Prospective	6.48	3.4	3.08	3.22	.001	.001
		Retrospective	11.00	3.4	7.61	7.85	<.001	< .001
Psychological scientists	Depression	Prospective	16.75	10.6	6.15	6.02	<.001	< .001
		Retrospective	20.26	10.6	9.66	11.22	<.001	< .001
	Life Satisfaction	Prospective	-15.13	1	16.13	-14.60	<.001	< .001
		Retrospective	-20.59	1	21.59	-22.24	<.001	< .001
	Implicit Prejudice	Prospective	14.00	-6.7	20.70	21.63	<.001	< .001
		Retrospective	9.82	-6.7	16.52	16.91	<.001	< .001
	Explicit Prejudice	Prospective	13.16	-123	136.16	152.93	<.001	< .001
		Retrospective	8.38	-123	131.38	122.66	<.001	< .001
	Loneliness	Prospective	15.83	-3.5	19.33	14.88	<.001	< .001
		Retrospective	24.87	-3.5	28.37	29.86	<.001	< .001

Political	Prospective	18.91	1	17.91	15.97	<.001	< .001
Polarization	Retrospective	25.92	1	24.92	24.31	<.001	< .001
Climate Change	Prospective	-2.10	-1.2	0.90	-0.96	.337	.359
	Retrospective	.59	-1.2	1.79	2.15	.032	.037
Traditionalism	Prospective	4.74	1.5	3.24	3.49	.001	< .001
	Retrospective	4.21	1.5	2.71	2.78	.006	.001
Violence	Prospective	1.09	26	24.91	-24.55	<.001	< .001
	Retrospective	2.94	26	23.07	-25.34	<.001	< .001
Charity	Prospective	3.33	1.5	1.83	1.55	.123	.137
	Retrospective	-.05	1.5	1.55	-1.47	.144	.157
	Prospective	-7.17	-1.3	5.87	-5.94	<.001	< .001
Generalized Trust	Retrospective	-16.92	-1.3	15.62	-17.64	<.001	< .001
Individualism	Prospective	2.96	3.4	0.44	-0.41	.684	.699
	Retrospective	6.99	3.4	3.59	4.14	<.001	< .001

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction.

Table S8. Regression models, prospective reports only (Studies 2a, 2b, 2c)

	Model 1 (null)		Model 2 (Group)		Model 3 (Domain)		Model 4 (Group + Domain)	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
(Intercept)	-0.009 (-0.049,0.032)	0.021	-0.044 (-0.107,0.018)	0.027	0.331 (0.172, 0.489)	0.081	0.326 (0.162,0.491)	0.084
Group [Scientist]	-	-	**0.061** (-0.021,0.143)**	0.042	-		0.01 (-0.092,0.111)	0.052
Domain (vs. Charitable Giving)								
Climate Change	-	-	-	-	**-0.923*** (-1.125, -0.721)**	0.103	**-0.925*** (-1.128,-0.722)**	0.103
Depression	-	-	-	-	**1.523*** (1.287, 1.758)**	0.12	**1.521*** (1.285, 1.757)**	0.12
Trust	-	-	-	-	-0.024 (-0.223, 0.176)	0.102	-0.025 (-0.226, 0.175)	0.102
Individualism	-	-	-	-	**-0.648*** (-0.848, -0.448)**	0.102	**-0.65*** (-0.85, -0.449)**	0.102
Life Satisfaction	-	-	-	-	**-1.867*** (-2.09, -1.644)**	0.114	**-1.869*** (-2.093, -1.645)**	0.114
Loneliness	-	-	-	-	**-1.869*** (-2.123, -1.615)**	0.129	**-1.869*** (-2.123, -1.615)**	0.129
Polarization	-	-	-	-	**0.73*** (0.521, 0.94)**	0.107	**0.728*** (0.518, 0.939)**	0.107
Traditionalism	-	-	-	-	**-0.441*** (-0.64, -0.241)**	0.102	**-0.442*** (-0.642, -0.242)**	0.102
Violence	-	-	-	-	**-0.224* (-0.445, -0.004)**	0.113	**-0.224* (-0.445, -0.004)**	0.112
R^2	0		0		0.231		0.231	
AIC	13056		13056		**11284**		11286	
BIC	13070		13077		**11362**		11372	
N	1101		1101		1101		1101	

Note. Significant effects indicated in bold. * $p < .05$, ** $p < .01$, *** $p < .001$

Table S9. Regression models, retrospective reports only (Studies 3a, 3b)

	Model 1 (null)		Model 2 (Group)		Model 3 (Domain)		Model 4 (Group + Domain)	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
(Intercept)	**0.125*** (0.077, 0.173)**	0.024	**0.157 (0.096, 0.219)**	0.027	0.009 (-0.146, 0.164)	0.079	0.061 (-0.102, 0.223)	0.083
Group [Scientist]	-	-	-0.082 (-0.18, 0.016)	0.05	-	-	**-0.132* (-0.261, -0.002)**	0.066
Domain (vs. Charitable Giving)								
Climate Change	-	-	-	-	**-0.876*** (-1.103, -0.65)**	0.115	**-0.876*** (-1.102, -0.65)**	0.115
Depression	-	-	-	-	**2.668*** (2.326, 3.009)**	0.174	**2.669*** (2.327, 3.01)**	0.174
Trust	-	-	-	-	**1.34*** (1.098, 1.582)**	0.124	**1.341*** (1.098, 1.583)**	0.124
Individualism	-	-	-	-	**0.494*** (0.275, 0.714)**	0.112	**0.495*** (0.275, 0.714)**	0.112
Life Satisfaction	-	-	-	-	**-2.369*** (-2.678, -2.06)**	0.157	**-2.369*** (-2.677, -2.06)**	0.157
Loneliness	-	-	-	-	**-2.997*** (-3.381, -2.613)**	0.196	**-2.997*** (-3.381, -2.613)**	0.196
Polarization	-	-	-	-	**2.57*** (2.239, 2.901)'**	0.169	**2.571*** (2.24, 2.902)**	0.169
Traditionalism	-	-	-	-	-0.081 (-0.297, 0.135)	0.11	-0.081 (-0.297, 0.135)	0.11
Violence	-	-	-	-	**0.455*** (0.235, 0.675)**	0.112	**0.455*** (0.235, 0.675)**	0.112
R^2	0		0		0.481		.482	
AIC	9313		9312		**6538**		6536	
BIC	9327		9332		**6613**		6618	
N	681		681		681		681	

Note. Significant effects indicated in bold. * $p < .05$, ** $p < .01$, *** $p < .001$

Table S10. Regression models for effect of judgment type (retrospective vs. prospective), Studies 2a-c, Studies 3a-b

	Model 1 (null)		Model 2 (Group)		Model 3 (Judgment + Domain)		Model 4 (Group + Domain)	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
(Intercept)	**0.163 (0.052, 0.273)****	0.056	**0.209*** (0.094, 0.323)**	0.059	0.044 (-0.073, 0.16)		0.078 (-0.046, 0.201)	0.063
Group [Scientist]	-	-	**-0.109** (-0.187, -0.031)**	0.04	-	-	-0.067 (-0.145, 0.011)	0.04
Judgment [Retrospective]	-	-	-	-	**0.245*** (0.167, 0.324)**		**0.233*** (0.154, 0.313)**	0.041
Domain (vs. Charitable Giving)								
Climate Change	**-0.86*** (-1.009, -0.712)**	0.076	**-0.847*** (-0.996, -0.698)**	0.076	**-0.837*** (-0.986, -0.688)**		**-0.83*** (-0.979**	0.076
Depression	**1.952*** (1.765, 2.138)**	0.095	**1.966*** (1.779, 2.152)**	0.095	**1.976*** (1.789, 2.163)**		**1.984*** (1.797, 2.171)**	0.096
Trust	**0.508*** (0.359, 0.657)**	0.076	**0.522*** (0.373, 0.67)**	0.076	**0.532*** (0.383, 0.681)**		**0.539*** (0.39, 0.689)**	0.076
Individualism	**-0.161* (-0.306, -0.016)**	0.074	**-0.148* (-0.293, -0.003)**	0.074	-0.137 (-0.283, 0.008)		-0.13 (-0.276, 0.015)	0.074
Life Satisfaction	**-1.965*** (-2.139, -1.792)**	0.089	**-1.952*** (-2.126, -1.779)**	0.089	**-1.942*** (-2.116, -1.768)**		**-1.935*** (-2.109, -1.761)**	0.089
Loneliness	**-2.24*** (-2.442, -2.038)**	0.103	**-2.24*** (-2.442, -2.038)**	0.103	**-2.241*** (-2.443, -2.039)**		**-2.24*** (-2.442, -2.038)**	0.103
Polarization	**1.326*** (1.162, 1.491)**	0.084	**1.34*** (1.176, 1.505)**	0.084	**1.351*** (1.187, 1.515)**		**1.359*** (1.194, 1.523)**	0.084
Traditionalism	**-0.258*** (-0.404, -0.113)**	0.074	**-0.245*** (-0.391, -0.1)**	0.074	**-0.235*** (-0.38, -0.09)**		**-0.228*** (-0.374, -0.082)**	0.074
Violence	0.118 (-0.038, 0.273)	0.079	0.117 (-0.038, 0.272)	0.079	0.117 (-0.038, 0.272)		0.117 (-0.039, 0.272)	0.079
R2	0.312		0.312		0.315		0.315	
AIC	18203		18197		18167		18166	
BIC	18288	-	18289	-	**18259**	-	18266	-

N	1782	-	1782	-	1782	-	1782	-

Note. Significant effects indicated in bold. * $p < .05$, ** $p < .01$, *** $p < .001$

Table S11. Accuracy rates of predictions as a function of degree psychological training (students, untenured faculty, tenured faculty)

Domain	Student	Untenured Faculty	Tenured Faculty	*p*	*p**
Life Satisfaction	0.17	0.12	0.11	0.77	0.94
Loneliness	0.16	0.22	0.16	0.98	0.98
Individualism	0.43	0.35	0.27	0.004	0.01
Charity	0.67	0.59	0.44	0.02	0.05
Climate Change	0.40	0.40	0.45	0.48	0.8
Traditionalism	0.50	0.48	0.47	0.83	0.94
Violence	0.52	0.50	0.55	0.85	0.94
Generalized Trust	0.68	0.59	0.57	0.05	0.1
Political Polarization	0.84	0.66	0.69	<.001	0.005
Depression	0.88	0.88	0.89	<.001	<.001

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction.

Table S12. Mean Difference and Significance Statistics for Comparisons between Psychology and Other Disciplines by Time in April 2020.

Time (months)	Domain	$M_{difference}$	*SE*	*z*	*p*	*p**
6	Explicit Prejudice	-2.32	2.28	-1.02	.309	.949
	Implicit Prejudice	-.17	2.23	-.08	.938	.949
	Individualism	-.30	2.24	-.14	.893	.949
	Traditionalism	6.57	2.24	2.93	.003	.112
	Generalized Trust	-2.55	2.22	-1.15	.250	.949
	Political Polarization	4.07	2.22	1.83	.067	.949
	Life Satisfaction	-2.46	2.20	-1.12	.265	.949
	Depression	1.44	2.22	.65	.517	.949
	Delay of Gratification	-.62	2.17	-.29	.775	.949
	Birth Rate	-1.40	2.20	-.64	.524	.949
	Climate Change	-.42	2.24	-.19	.850	.949
12	Explicit Prejudice	-.97	2.28	-.43	.670	.949
	Implicit Prejudice	.73	2.23	.33	.742	.949
	Individualism	-.35	2.24	-.16	.876	.949
	Traditionalism	2.91	2.24	1.30	.195	.949
	Generalized Trust	-2.71	2.24	-1.21	.227	.949
	Political Polarization	2.74	2.22	1.23	.218	.949
	Life Satisfaction	-1.11	2.20	-.50	.615	.949
	Depression	-.76	2.22	-.34	.731	.949
	Delay of Gratification	2.03	2.17	.94	.350	.949
	Birth Rate	.26	2.20	.12	.907	.949
	Climate Change	1.87	2.24	.83	.404	.949
24	Explicit Prejudice	-1.89	2.28	-.83	.406	.949
	Implicit Prejudice	1.63	2.23	.73	.465	.949
	Individualism	-2.38	2.24	-1.06	.289	.949
	Traditionalism	1.23	2.24	.55	.583	.949
	Generalized Trust	-1.56	2.24	-.69	.487	.949
	Political Polarization	.59	2.24	.26	.792	.949
	Life Satisfaction	3.15	2.20	1.43	.153	.949
	Depression	.14	2.24	.06	.949	.949
	Delay of Gratification	.20	2.17	.09	.927	.949
	Birth Rate	1.08	2.20	.49	.623	.949
	Climate Change	2.09	2.24	.94	.350	.949

Note. *P*-values in the rightmost column were adjusted using the Benjamini-Hochberg false discovery rate correction. $M_{difference}$ refers to a difference score between mean of experts from psychology and mean of experts from other disciplines.

Table S13. Mean Difference and Significance Statistics for Comparisons between Psychology and Other Disciplines by Time in May 2020.

					Psychology vs. non-Psychology			
Time	Domain	*M*	*SE*	*95% CI**	$M_{difference}$	*z*	*p*	*p**
6 months	Explicit Prejudice	13.04	.91	1.78	0.45	0.20	.842	.975
	Implicit Prejudice	13.83	.91	1.79	-1.89	-0.84	.404	.699
	Individualism	2.83	.92	1.80	0.96	0.42	.673	.842
	Traditionalism	4.64	.92	1.80	-6.10	-2.67	.008	.131
	Generalized Trust	-7.19	.92	1.80	-4.91	-2.16	.031	.230
	Political Polarization	18.86	.91	1.79	1.25	0.56	.576	.810
	Life Satisfaction	-15.20	.91	1.79	3.38	1.48	.138	.444
	Depression	16.72	.92	1.80	0.77	0.34	.733	.892
	Delay of Gratification	.01	.92	1.80	-5.49	-2.44	.015	.131
	Birth Rate	1.04	.92	1.80	0.09	0.04	.969	.991
	Violence	.96	.97	1.89	-5.82	-2.45	.014	.131
	Religiosity	7.13	.96	1.88	-1.19	-0.50	.617	.817
	Loneliness	15.72	.95	1.87	-4.84	-2.04	.041	.265
	Charity	3.21	.95	1.87	0.33	0.14	.889	.975
	Climate Change	-2.20	.92	1.80	-1.81	-0.80	.426	.711
12 months	Explicit Prejudice	9.19	.91	1.78	-0.09	-0.04	.967	.991
	Implicit Prejudice	10.62	.91	1.79	-1.90	-0.84	.403	.699
	Individualism	3.36	.92	1.80	0.36	0.16	.874	.975
	Traditionalism	2.93	.92	1.80	-6.14	-2.69	.007	.131
	Generalized Trust	-4.68	.92	1.80	-3.26	-1.44	.151	.453
	Political Polarization	14.58	.91	1.79	1.89	0.84	.399	.699
	Life Satisfaction	-10.03	.91	1.79	4.07	1.78	.075	.336
	Depression	13.43	.92	1.80	-0.24	-0.11	.914	.980
	Delay of Gratification	-1.30	.92	1.80	-4.07	-1.81	.070	.336
	Birth Rate	2.26	.92	1.80	2.62	1.15	.252	.566
	Violence	5.51	.97	1.89	-3.56	-1.50	.134	.444
	Religiosity	4.02	.96	1.88	-2.08	-0.88	.381	.699
	Loneliness	7.99	.95	1.87	-4.32	-1.82	.069	.336
	Charity	.38	.95	1.87	0.37	0.16	.874	.975
	Climate Change	3.39	.91	1.79	-2.72	-1.20	.231	.566
24 months	Explicit Prejudice	6.03	.91	1.78	-1.01	-0.44	.657	.842
	Implicit Prejudice	7.40	.91	1.79	-1.74	-0.77	.443	.712
	Individualism	3.78	.92	1.80	1.64	0.72	.469	.728
	Traditionalism	.83	.92	1.81	-5.63	-2.46	.014	.131
	Generalized Trust	-3.05	.92	1.80	-2.46	-1.09	.278	.595
	Political Polarization	11.94	.91	1.79	2.82	1.26	.209	.554
	Life Satisfaction	-3.07	.91	1.79	3.41	1.50	.134	.444
	Depression	8.09	.92	1.80	-1.52	-0.67	.503	.755
	Delay of Gratification	-1.38	.92	1.80	-1.20	-0.53	.594	.810
	Birth Rate	-1.38	.92	1.80	1.24	0.54	.588	.810
	Violence	5.44	.97	1.89	-2.74	-1.15	.249	.566
	Religiosity	1.61	.96	1.88	-2.98	-1.26	.209	.554
	Loneliness	2.06	.95	1.87	-4.06	-1.71	.087	.355
	Charity	-.29	.95	1.87	-0.02	-0.01	.993	.993
	Climate Change	8.91	.91	1.79	-1.97	-0.87	.384	.699

Note. CI indicates 95% confidence distance from the mean. Right column indicates *p*-values adjusted for number of tests using Benjamini-Hochberg false discovery rate procedure. $M_{difference}$ refers to a difference score between mean of experts in Psychology and other disciplines.

Table S14. Comparisons of Retrospective Estimates between Psychological Scientists and Lay People in October/November 2020 using Frequentist Methods.

Dimension	$M_{difference}$	*SE*	*z*	*p*	*p**
Explicit Prejudice	-1.66	1.40	-1.18	.238	.297
Implicit Prejudice	-2.59	1.41	-1.84	.066	.123
Individualism	-4.02	1.40	-2.87	.004	.015
Traditionalism	-.25	1.40	-.18	.859	.859
Generalized Trust	-4.41	1.40	-3.15	.002	.008
Political Polarization	-2.05	1.40	-1.46	.144	.216
Life Satisfaction	-1.77	1.40	-1.26	.207	.283
Depression	-3.81	1.40	-2.72	.006	.019
Delay of Gratification	-5.56	1.40	-3.97	< .001	.001
Birth Rate	-2.28	1.40	-1.63	.104	.173
Violence	-8.01	1.41	-5.66	< .001	< .001
Religiosity	-1.50	1.41	-1.06	.289	.332
Loneliness	-3.17	1.41	-2.24	.025	.062
Charity	-1.44	1.41	-1.02	.310	.332
Climate Change	-2.73	1.40	-1.95	.051	.110

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction.

Table S15. Comparisons of Retrospective Estimates between Psychological scientists and Lay People in October/November 2020 using Bayesian Statistics.

Dimension	*β*	*Estimated Error*	*95%CI (Lower)*	*95%CI (Upper)*	*CI Includes Zero*	*Bayes Factor*	*Strength of Evidence*
Traditionalism	.01	.08	-.14	.17	Yes	62.61	Very Strong (H0)
Charity	.07	.08	-.08	.23	Yes	43.33	Very Strong (H0)
Explicit Prejudice	.08	.08	-.07	.24	Yes	36.21	Very Strong (H0)
Life Satisfaction	.10	.08	-.06	.25	Yes	31.00	Very Strong (H0)
Religiosity	.10	.08	-.05	.26	Yes	26.61	Strong (H0)
Political Polarization	.11	.08	-.04	.26	Yes	25.72	Strong (H0)
Implicit Prejudice	.14	.08	-.02	.29	Yes	13.48	Strong (H0)
Climate Change	.15	.08	.00	.31	No	9.48	Moderate (H0)
Birth Rate	.18	.08	.03	.34	No	4.29	Moderate (H0)
Loneliness	.18	.08	.03	.33	No	4.21	Moderate (H0)
Individualism	.23	.08	.07	.38	No	1.18	Anecdotal (H0)
Generalized Trust	.24	.08	.08	.39	No	.80	Anecdotal (H1)
Depression	.24	.08	.08	.39	No	.47	Anecdotal (H1)
Delay of Gratification	.29	.08	.12	.45	No	.26	Moderate (H1)
Violence	.43	.08	.27	.59	No	< 0.01	Extreme (H1)

Note: 95% CI refers to Bayesian credible interval and Bayes Factor denotes BF01. Bayes Factor interpretation is based on Lee & Wagenmakers, 2013.

Table S16. Comparisons of Psychological Scientists' Retrospective Estimates in October/November 2020 by Self-Reported Expertise in Domain of Forecast.

Dimension	$M_{difference}$	*SE*	*z*	*p*	*p**
Explicit Prejudice	4.71	1.88	2.50	.012	.186
Implicit Prejudice	3.48	1.82	1.92	.055	.225
Individualism	3.86	2.05	1.88	.060	.225
Traditionalism	.93	2.49	.37	.709	.792
Generalized Trust	1.56	2.24	.70	.485	.727
Political Polarization	-1.79	2.09	-.86	.393	.727
Life Satisfaction	1.70	1.84	.92	.356	.727
Depression	.81	1.88	.43	.666	.792
Delay of Gratification	-.86	1.94	-.44	.658	.792
Birth Rate	-1.51	4.52	-.33	.739	.792
Violence	-6.05	3.42	-1.77	.077	.230
Religiosity	-.21	2.64	-.08	.937	.937
Loneliness	-1.54	2.01	-.77	.442	.727
Charity	2.82	3.09	.91	.361	.727
Climate Change	5.65	2.98	1.90	.058	.225

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction.

Table S17. Comparisons of the Effects of News Reports and Vivid Memories on Retrospective Estimates by Sample and Dimension.

Dimension	Type	$M_{difference}$	*SE*	*z*	*p*	*p**	*Extremeness*
Lay People							
Birth Rate	News	3.27	2.24	1.46	.144	.188	Unchanged
	Vivid	.11	3.66	.03	.975	.975	Unchanged
Charity	News	8.67	2.05	4.22	<.001	<.001	Greater
	Vivid	7.15	2.18	3.28	.001	.003	Greater
Climate Change	News	14.09	1.68	8.39	<.001	<.001	Greater
	Vivid	9.85	2.00	4.93	<.001	<.001	Greater
Delay of Gratification	News	1.95	2.13	.92	.360	.400	Unchanged
	Vivid	1.65	1.85	.89	.372	.436	Unchanged
Depression	News	5.62	1.74	3.24	.001	.003	Greater
	Vivid	5.82	1.70	3.42	.001	.002	Greater
Explicit Prejudice	News	8.28	1.70	4.87	<.001	<.001	Greater
	Vivid	8.06	1.95	4.13	<.001	<.001	Greater
Generalized Trust	News	-2.61	1.77	-1.47	.142	.188	Unchanged
	Vivid	-4.54	1.70	-2.67	.008	.015	Unchanged
Implicit Prejudice	News	9.66	1.71	5.63	<.001	<.001	Greater
	Vivid	5.57	1.85	3.01	.003	.006	Unchanged
Individualism	News	6.97	1.79	3.89	<.001	<.001	Greater
	Vivid	1.72	1.96	.88	.378	.436	Unchanged
Life Satisfaction	News	-.17	1.77	-.10	.924	.924	Unchanged
	Vivid	-2.68	1.70	-1.58	.114	.163	Unchanged
Loneliness	News	7.83	1.69	4.62	<.001	<.001	Greater
	Vivid	7.21	1.70	4.24	<.001	<.001	Greater
Political Polarization	News	13.00	1.97	6.59	<.001	<.001	Greater
	Vivid	4.43	1.70	2.61	.009	.016	Unchanged
Religiosity	News	7.59	2.31	3.28	.001	.003	Greater
	Vivid	7.60	2.14	3.55	<.001	.002	Greater
Traditionalism	News	3.78	1.94	1.95	.051	.081	Unchanged
	Vivid	5.42	2.05	2.64	.008	.015	Unchanged
Violence	News	12.95	1.73	7.49	<.001	<.001	Greater
	Vivid	11.55	2.22	5.21	<.001	<.001	Greater
Psychological scientist							
Birth Rate	News	-.93	2.66	-.35	.728	.753	Unchanged
	Vivid	-4.14	3.36	-1.23	.218	.275	Unchanged
Charity	News	3.89	2.34	1.66	.096	.138	Unchanged
	Vivid	7.87	2.25	3.50	<.001	.002	Greater
Climate Change	News	7.87	2.07	3.80	<.001	<.001	Greater
	Vivid	9.09	2.42	3.76	<.001	.001	Greater
Delay of Gratification	News	3.13	2.57	1.22	.223	.279	Unchanged
	Vivid	-1.44	2.37	-.61	.544	.583	Unchanged
Depression	News	4.04	2.18	1.85	.064	.096	Unchanged
	Vivid	7.32	2.16	3.39	.001	.002	Greater
Explicit Prejudice	News	2.32	2.39	.97	.331	.382	Unchanged
	Vivid	5.28	2.24	2.36	.018	.030	Unchanged
Generalized Trust	News	-4.95	2.11	-2.35	.019	.038	Unchanged
	Vivid	-9.13	2.14	-4.27	<.001	<.001	Greater
Implicit Prejudice	News	5.99	2.09	2.87	.004	.009	Unchanged
	Vivid	3.52	2.43	1.45	.147	.200	Unchanged
Individualism	News	4.25	2.12	2.00	.045	.080	Unchanged
	Vivid	1.67	2.47	.67	.500	.556	Unchanged
Life Satisfaction	News	-4.26	2.15	-1.98	.048	.080	Unchanged
	Vivid	-4.10	2.09	-1.96	.050	.075	Unchanged
Loneliness	News	5.75	2.17	2.65	.008	.017	Unchanged
	Vivid	6.45	2.12	3.05	.002	.005	Unchanged
Political Polarization	News	2.83	2.66	1.06	.288	.345	Unchanged
	Vivid	6.56	2.09	3.14	.002	.004	Greater
Religiosity	News	1.62	3.12	.52	.604	.647	Unchanged
	Vivid	.86	3.80	.23	.821	.849	Unchanged
Traditionalism	News	5.21	2.24	2.33	.020	.038	Unchanged
	Vivid	3.52	2.87	1.23	.220	.275	Unchanged
Violence	News	6.56	2.14	3.06	.002	.005	Greater
	Vivid	6.11	2.93	2.08	.037	.059	Unchanged

Note: **p* = FDR adjusted values. $M_{difference}$ = subtracting the mean of non-vivid from vivid and non-concrete news from concrete-news. Extremeness = are vivid and concrete-news estimates lower (absolute value of estimates closer to baseline), greater (absolute value of estimates farther from baseline) or unchanged.

Table S18. Comparisons of Prospective and Retrospective Estimates by Sample and Dimension.

Sample	Dimension	$M_{difference}$	*SE*	*z*	*p*	*p**	*Extremeness*
Lay People	Birth Rate	3.73	1.39	2.69	.007	.011	Lower
	Charity	4.18	1.39	3.00	.003	.005	Lower
	Climate Change	-1.82	1.39	-1.32	.188	.211	Unchanged
	Delay of Gratification	-.46	1.39	-.33	.739	.739	Unchanged
	Depression	-7.33	1.39	-5.29	< .001	< .001	Greater
	Explicit Prejudice	-1.79	1.39	-1.21	.197	.211	Unchanged
	Generalized Trust	7.76	1.39	5.60	< .001	< .001	Greater
	Implicit Prejudice	-4.54	1.39	-3.28	.001	.002	Greater
	Individualism	-4.54	1.39	-3.27	.001	.002	Greater
	Life Satisfaction	9.32	1.39	6.72	< .001	< .001	Greater
	Loneliness	-11.85	1.39	-8.51	< .001	< .001	Greater
	Political Polarization	-11.36	1.39	-8.18	< .001	< .001	Greater
	Religiosity	2.28	1.39	1.63	.102	.128	Unchanged
	Traditionalism	-2.36	1.39	-1.70	.088	.121	Unchanged
	Violence	-8.34	1.39	-5.99	< .001	< .001	Greater
Psychological scientists	Birth Rate	.97	1.19	.82	.415	.445	Unchanged
	Charity	2.90	1.41	2.05	.040	.050	Lower
	Climate Change	-3.59	1.18	-3.03	.002	.005	Lower
	Delay of Gratification	4.34	1.19	3.65	< .001	.001	Greater
	Depression	-3.16	1.19	-2.66	.008	.011	Greater
	Explicit Prejudice	5.02	1.19	4.22	< .001	< .001	Lower
	Generalized Trust	10.51	1.19	8.86	< .001	< .001	Greater
	Implicit Prejudice	3.54	1.19	2.96	.003	.005	Lower
	Individualism	-7.92	1.19	-6.67	< .001	< .001	Greater
	Life Satisfaction	5.08	1.19	4.28	< .001	< .001	Greater
	Loneliness	-9.52	1.41	-6.73	< .001	< .001	Greater
	Political Polarization	-10.88	1.19	-9.16	< .001	< .001	Greater
	Religiosity	4.06	1.42	2.86	.004	.006	Lower
	Traditionalism	.22	1.19	.18	.857	.857	Unchanged
	Violence	-2.34	1.43	-1.64	.101	.116	Unchanged

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction. $M_{difference}$ was computed by subtracting the mean of retrospective from prospective estimates. Extremeness signifies whether retrospective estimates were: lower (absolute value of retrospective estimates closer to baseline than prospective), greater (absolute value of retrospective estimates farther away from baseline than prospective) or unchanged (no difference between the two types of estimates) as compared to prospective estimates.

Table S19. Comparisons of Prospective and Retrospective Estimates by Sample and Dimension Controlling for Demographics.

Sample	Dimension	$M_{difference}$	*SE*	*z*	*p*	*p**
Lay People	Birth Rate	3.68	1.39	2.65	.008	.012
	Charity	4.18	1.39	3.00	.003	.004
	Climate Change	-1.91	1.39	-1.38	.169	.195
	Delay of Gratification	-.49	1.39	-.36	.723	.723
	Depression	-7.33	1.39	-5.29	< .001	< .001
	Explicit Prejudice	-1.75	1.39	-1.26	.207	.222
	Generalized Trust	7.71	1.39	5.56	< .001	< .001
	Implicit Prejudice	-4.56	1.39	-3.29	.001	.002
	Individualism	-4.59	1.39	-3.31	.001	.002
	Life Satisfaction	9.28	1.39	6.68	< .001	< .001
	Loneliness	-11.87	1.39	-8.52	< .001	< .001
	Political Polarization	-11.35	1.39	-8.17	< .001	< .001
	Religiosity	2.30	1.40	1.65	.099	.124
	Traditionalism	-2.35	1.39	-1.70	.090	.123
	Violence	-8.40	1.39	-6.02	< .001	< .001
Psychological scientists	Birth Rate	2.23	1.40	1.60	.111	.128
	Charity	3.34	1.43	2.34	.019	.027
	Climate Change	-2.77	1.39	-1.99	.047	.058
	Delay of Gratification	3.65	1.40	2.62	.009	.017
	Depression	-3.41	1.39	-2.45	.014	.021
	Explicit Prejudice	4.75	1.39	3.41	.001	.002
	Generalized Trust	10.29	1.39	7.39	< .001	< .001
	Implicit Prejudice	4.10	1.40	2.94	.003	.007
	Individualism	-3.45	1.39	-2.48	.013	.021
	Life Satisfaction	6.19	1.39	4.45	< .001	< .001
	Loneliness	-9.40	1.43	-6.58	< .001	< .001
	Political Polarization	-6.81	1.39	-4.90	< .001	< .001
	Religiosity	4.53	1.43	3.16	.002	.004
	Traditionalism	.73	1.40	.52	.601	.601
	Violence	-1.57	1.44	-1.09	.276	.296

Note: Rightmost *p* value column was adjusted for false discovery rate using Benjamini-Hochberg correction. $M_{difference}$ was computed by subtracting the mean of retrospective from prospective estimates. Extremeness signifies whether retrospective estimates were: lower (absolute value of retrospective estimates closer to baseline than prospective), greater (absolute value of retrospective estimates farther away from baseline than prospective) or unchanged (no difference between the two types of estimates) as compared to prospective estimate.