nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Insights into accuracy of social scientists' forecasts of societal change
Corresponding author name(s): Igor Grossmann

Reviewer Comments & Decisions:

Decision Letter, initial version:

15th August 2022

Dear Professor Grossmann,

Thank you once again for your manuscript, entitled "Insights into accuracy of social scientists' forecasts of societal change", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We ask you to specifically address the several fundamental methodological concerns the reviewers raise, providing also full documentation and code for your analyses, as well as data to enable thorough evaluation.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Samantha Antusch

Samantha Antusch, PhD
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: large-scale social science experiments ; social science ; open science

Reviewer #2: large-scale social science experiments ; social science

Reviewer #3: expert forecasting

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

Overall, this is an impressive collaboration that yields important data about the readiness (or lack thereof) of social scientists to predict social change and/or influence policy. A major strength is the comparison of the social scientists' accuracy to both naïve statistical models and a comparison sample of non-experts from Prolific, which helps put the (in)accuracy in context. That the authors also include well documented open data and analysis code is a testament to their diligent work on the project.

The conclusion that social scientists may not be very accurate at predicting societal change would not be a great surprise to me. However, there are some aspects of the results that I think are worth considering. Some of the tables and figures make me wonder whether the scientists may have done better than it seems:

-Averaging across all 12 domains (Table 1), scientists outperformed the lay crowd by Cohen's d = 0.492 on average (this is a larger effect size than I would have predicted in advance, and much larger than I expected given the results of the inferential tests). They performed better (ignoring significance/bayes) in 11/12 domains. With 86 teams/359 forecasts, I feel like there's a disconnect between these effect sizes and the non-significant inferential tests (and indeed the Bayes factors signaling strongly in favor of no effect for the most part). Am I missing something? Maybe the Cohen's d is sort of misleading because it's computed on repeated measures?
Relatedly, it would help to clarify exactly what analysis is being run on P10 "Comparing accuracy of scientific and lay forecasts, we observed no significant difference between social scientist and lay crowds, $F(11, 1747) = 0.88$, $P = .349$, $R2 < 0.001$" (this may also apply to other reported results, I would clarify the analyses being run)

-Figure 1 plots only the best performing naïve statistical benchmark per domain. Considering the variability in performance between these benchmarks by domain (Figure 2), is this a fairer comparison than plotting the mean or median benchmark performance?

I'd make a similar point about the results at the top of P14. Having to beat all three benchmarks seems like a tall order given the seemingly random variability in performance between the benchmarks in any given domain (although perhaps there's an argument that the scientists should be easily beating all three). Overall, these results left me with the impression that the scientists were consistently underperforming the benchmarks, when Fig S2 seems to show scientists being competitive (or better than) the benchmarks quite regularly. I think it might be more informative to either come up with some composite of the benchmarks to compare against, or report performance of scientists vs each benchmark separately.

-MASE was a new metric to me (and I'm not an expert in any of these naïve forecasting benchmarks), so I appreciated the guidelines for interpreting it. To be clear, is it correct that a MASE < 1 (e.g. in Figure 1) would be quite impressive in the current context? (given that 1.38 is considered "good" prediction in tourism, and 1.76 being better than median data science competitions). I assume this is because the one-step random walk is benefitting from being in-sample, vs the out-of-sample predictions provided for the whole year from scientists? If I'm understanding that correctly, I think it may be worth mentioning something along those lines as a caveat in the results reporting. On my initial pass I got the impression that not hitting MASE < 1.0 would be quite meaningful, but after digging in to the calculation a bit more I'm thinking it's actually quite a high bar in this context.

Of course, it's entirely possible that there are good responses to each of these points, but I'd like the authors to consider them and make sure the conclusions are well calibrated to the data. Then, more minor suggestions that some readers may find helpful:

-From the text, it's not clear to me what tests/comparisons were pre-registered and which were exploratory. Please clarify this.

-On page 9, it may help to clarify that participants were predicting future values for the same exact polls/data that they were provided as historical data. (this is obvious in the supplemental materials, but wasn't obvious to me from the text)

-On P10, the results in the first paragraph (comparing social scientists vs in-sample random walks and prior forecasting competitions) are reported descriptively, whereas the next paragraph (comparing social scientists to lay crowd) rely on inferential statistics. Is this as intended?

-Potentially useful citation with some counter-points to the Van Bavel et al. paper [47], and echo some of the points in the current discussion:
IJzerman, H., Lewis, N.A., Przybylski, A.K. et al. Use caution when applying behavioural science to policy. Nat Hum Behav 4, 1092–1094 (2020). https://doi.org/10.1038/s41562-020-00990-w

(also note that the Van Bavel paper, and some others, appear only in the references list and are not cited in the text)

-Table 1 seems to be missing the 95% CI for the d scores.

Thanks to the authors for this massive effort.

Rick Klein

Reviewer #2:
Remarks to the Author:

In this manuscript the authors conducted two forecasting tournaments where social scientists where ask to predict future of aggregated rates in the United States (e.g., political polarization as measured by difference in presidential approval ratings between Democrats and Republicans). These "expert" predictions were compared to a variety of benchmarks, some statistical (e.g., random walk) and a wisdom-of-the-crowd approach. The authors find that statistical models generally do better than experts who generally do better than the crowd (Fig 1). The authors perform several additional analyses, such as comparing accuracy across domains, accuracy across tournaments, and accuracy by strategy.

Overall, I want to thank the authors for this interesting paper, and I want to especially thank them for doing a mass collaboration. Team science is challenging for many reasons, but it can lead to really interesting research, such as this paper.

I have several questions/concerns/suggestions/provocations that I hope will help the authors improve the paper. These are grouped in major (in the sense that I really think they should be addressed) and minor (in the sense that I think they may be addressed or not, as the authors and editors see fit). They are listed in no order.

Major:

- Heterogeneity within participants. One of the main ways that this differs from other projects using the common task method is that the authors don't simply present results from the best forecaster, they present average results. This means that a small number of bad forecasters could pull down the average performance of the experts. I would like to understand more about the heterogeneity of responses. Are a small group of bad "experts" pulling down the overall score? Perhaps this is represented in Figure 1, but it is not clear to me what the intervals are around the "scientists" predictions. Perhaps something like Fig 3 could include the empirical distribution of errors (not just the standard deviation) or could be presented in terms of median. Naturally, I would leave it to the authors to come up with the way they think is best to address this concern.

- Timing. It is not clear to me exactly which months are being forecast. The start dates are shown on page 9 lines 276-277. This should be clearer. Also, for future readers and readers from outside the United States, I think it should be clearer that this was perhaps an unusual time in the US due to COVID and the Black Lives Matter protests. Perhaps some kind of timeline figure could be added to the SM, which would also include key historical events happening during these periods (e.g., George Floyd murder, COVID vaccine rollout, etc)? Also, I think the authors have the capacity to check empirically whether their outcomes moved in unusual ways during this time.

- Benchmarks. I thank the authors for thinking carefully about benchmarks. This is important and unfortunately relatively rare. When constructing the historical mean benchmark, I don't understand why the authors didn't use the historical mean (they used random resampling instead). Also, in Figure 1 and Figure 4, I don't understand why the M4 tournament performance is a reasonable benchmark. To be clear, I'm not saying that it is not, but I'm not familiar with the M4 tournament, and the little I do know makes it seem quite different.

- Construct validity of outcomes: Some of the outcomes are complex measurements on non-random samples. For example, affective well-being is a complex summary of Twitter. The authors seem aware of these issues, but I wonder how much of the forecasting challenges are created by measurement challenges.

- I am very surprised that accuracy was better for predictions further the future (line 413). I also didn't understand the proposed explanation in the text and Fig S1. Do the authors think this will be a general pattern in other forecasting tournaments or might it be specific to this time period, which the authors acknowledge is unusual? Also, do the authors have other evidence to support their anchoring hypothesis from the crowd or from the month-by-month predictions? Relatedly, in the next paragraph, the authors seem to conclude that forecast accuracy is a function of the amount of timeseries data available. I find it surprising that six months of additional data would make much difference, given that they already had 39 months of data (if I understand the design correctly). If this is true, it suggests that the length of the historical data window is an important design consideration, and then I

would wonder why the authors picked the historical data window they did.

- Discussion: I wonder if the authors have any new theories about which domains are more difficult to forecast based on these results. Based on around line 527 the authors seem to think that this is mainly related to historical volatility. Could this be tested directly? Also, if this is only about volatility that would be important to know so that people don't try to develop theories based on the substantive characteristics of these outcomes.

- I really like the idea of Fig S1. In some outcomes the ground truth appears to be a natural continuation of the historical timeseries. But for other outcomes like "Exp Bias vs Asian Americans" and "Exp Bias vs African Americans" there appears to be a kind of discontinuous jump. Perhaps I'm reading too much into just two points in the historical data. Do the authors have any ideas about whether there might be sharp changes in some outcomes right around the tournament time and if so why this might be the case?

Minor:

- I believe there is a typo on line 1165 (could be 5,000 – 6,000 or 50,000 to 60,000).

- The term "expert" is used quite loosely here. Do we think that graduate students are "experts" in social sciences? To be clear, I think it is great to assess the accuracy of graduate students, I'm just expressing concern about the construct validity of the term "experts" given the pool of participants.

- The design that these researchers used was unusual, in part because there are not yet clear standards in this area. I wonder if the authors should reflect on what they learned about designing forecasting tournaments and include that in the article. For example, it seems like participants were able to choose what domains they forecasted, and this probably made the analysis more complex. Do the authors think this complexity was worth it? Also, the authors had two nested tournaments. Is that something they think must be done in the future as well? How would they recommend that future researchers select domains?

- I was not able to find the pre-registered analytic plan on Open Science Framework (OSF). I'm sure that this represents an error on my part; I am not familiar with this website. However, I would hope that it can be reviewed. Perhaps in the future the authors could provide a direct link to the document with the historical timestamp for reviewers and readers who are not as familiar with OSF.

- The authors write "Fig 1 shows that in Tournament 1, social scientists' forecasts were, on average, inferior to in-sample random walks in all domains." (line 322). However, it seems that for some domains the marker for "scientists" is to the left of the market for "naïve statistic". I hope that the match between the text and figure could be clarified.

- Several places around line 323 the authors talk about a "in-sample random walk", and I'm not sure what that is or how that compares the rank walk described on line 315.

- I found it very interesting that there was little overlap between the winners in tournament 1 and tournament 2. However, focusing just on the winners discards a lot of information. Is it the case that teams that did well in tournament 1 did well in tournament 2? If there is a lot of reversion to the

mean that would suggest that luck partially explains some of the good performance in tournament 1.

- On line 332, I don't understand what this R2 value of less than 0.001 represents.

- I didn't understand the claims on line 336 about Bayesian analysis, and it was not obvious to me where to find the necessary details in the SM.

- Around like 430 I didn't understand what is meant by model accuracy. Is model being used with interchangeably with team?

- On line 597 the authors suggest that social scientists could benefit from testing whether a trend is stochastic or deterministic. I agree that this would be very valuable knowledge, but it seems impossible to test (at least to me). If the authors know how to do this, it would be great if they could share a bit more. If this is impossible, then perhaps it should be removed?

- On line 606 the authors write that the ability to accurately predict trends in these variables "would appear to be of critical importance." It is not clear to me that predicting these trends slightly more accurately than a naïve model is really important. Did the authors include "appear" here as a hedge? Could they explain more about why slightly better predictions would be "of critical importance". Also, just to be clear, I'm not saying that the domains they studied are not important, just that slightly better prediction of those domains is not obviously important.

- I really like the idea of Fig S1, but I find it hard to interpret. In the negative affect panel, for example, why does it seem that there are vertical blue bands starting in November?

I wanted to thank the authors again for an interesting, important, and stimulating article. I think it has the potential to be read by a wide range of scholars, and I hope that the feedback provided above helps make the manuscript clearer and ultimately more impactful.

Reviewer #3:
Remarks to the Author:

Review Nature 20550_0_art_file_3920849_rdhjnm

The purpose of this ms is captured in the following statement:

Line 236-244: "Prior forecasting initiatives have not fully addressed this question [predicting trends in social phenomena] for two primary reasons. First, forecasting initiatives with subject matter experts have focused on examining the probability of occurrence for specific one-time events (4, 6) rather than the accuracy of ex-ante predictions of societal change over multiple units of time (7). The likelihood of a prediction regarding a one-off event being accurate is higher than that of a prediction regarding societal change across an extended time period. Second, forecasting efforts have concentrated on predicting geopolitical (4) or economic events (8) rather than broader societal phenomena. "

An enormous amount of material is described. A proper review must check the integrity of the data, the analytical methods and the conclusions. I cannot conduct such a review because the authors do not provide nearly enough material. I found forecast data for, presumably, Tournament 1 (https://osf.io/6wgbj/), but I found no forecast data for Tournament 2. A proper review must access all the data behind Fig S1, S2. I miss significant details for computing the MASE (what is the "training MAE" data). I cannot understand or reproduce Table 1 as the lay forecasts and realizations are not given. I cannot determine whether the statistical tests are appropriate. Etc etc.

A variance decomposition of the Tournament 1 forecast data raised concerns regarding the suitability of this data for the purpose of the study as quoted above. There are 88 teams in total, each assessing a maximum of 12 issues over a period of 12 months. Three of the teams are "revised" (1859revised, BlackSwanrevised and R4VST9 revised). That gives 85 unrevised teams, not 86 (line 204). Counting revisions there are 363 forecasts. Excluding revisions there are 363 – 7×12 = 279 not 359 (line 204). Am I looking at the right data? What am I missing?

The number of teams assessing each of the 12 issues are shown below, ranging from 21 to 58.

It is evident that the various issues are assessed by different teams, employing different methods. Does team performance take the number of assessments into account? Are teams with few assessments choosing the better part of valor?

My concern with this data relative to the research question is that the forecasts are rather insensitive to time, as are many of the issues. For each issue, there is one forecast per team per month for 12 months. Below is the table for the 23 team assessments of Explicit African American Bias (eafric). Column and row averages and variances are shown. The variance in the 23×12 table equals the variance of the average + the average of the variance, computed either column- or row-wise = 0.003295764. The variances row-wise are very small, 20 of the 23 teams are below 0.001, nearly half are below 0.0001. This means the individual teams show little month-to-month variation. On the other hand, the column variances are all above 0.001; after the first two months all are nearly equal to the total variance. The Teams explain 99.54% of the variance in this data, the Months explain 13.04% (if teams and months were independent these numbers would add to 100%). Put simply, choosing a month does not reduce the variation in forecasts, but choosing a Team does.

Similar patters hold for many of the other 11 issues:

Using population weighted means and variances, the issue explains 99.44% of the overall variance.

My concern is that this data is unsuitable for answering the question whether social scientists can predict trends in social phenomena. As the authors note: "for half of the domains the average forecasts were highly similar or nearly indistinguishable from the last historical data points provided to forecasting teams (Fig. S2)" Don't you mean fig S1?? (line 1614.). No information is provided how the

values in Figure S1 were computed (stereo-consistent?). There may be an important message in this data but the reader can't extract without much more information.

Author Rebuttal to Initial comments

Response to **Reviewer #1**

Overall, this is an impressive collaboration that yields important data about the readiness (or lack thereof) of social scientists to predict social change and/or influence policy. A major strength is the comparison of the social scientists' accuracy to both naïve statistical models and a comparison sample of non-experts from Prolific, which helps put the (in)accuracy in context. That the authors also include well documented open data and analysis code is a testament to their diligent work on the project.

Thank you!

The conclusion that social scientists may not be very accurate at predicting societal change would not be a great surprise to me. However, there are some aspects of the results that I think are worth considering. Some of the tables and figures make me wonder whether the scientists may have done better than it seems:

Averaging across all 12 domains (Table 1), scientists outperformed the lay crowd by Cohen's d = 0.492 on average (this is a larger effect size than I would have predicted in advance, and much larger than I expected given the results of the inferential tests). They performed better (ignoring significance/bayes) in 11/12 domains. With eighty-six teams/359 forecasts, I feel like there's a disconnect between these effect sizes and the non-significant inferential tests (and indeed the Bayes factors signaling strongly in favor of no effect for the most part). Am I missing something? Maybe the Cohen's d is sort of misleading because it's computed on repeated measures?

Thank you for raising this issue, which we also discussed at length among the members of the Forecasting Collaborative (our discussion resulted in the section on "Multiverse analyses of domain-general accuracy" in the Supplementary materials, where we compare and review different approaches). The apparent disconnect between p-values and Bayes factor scores on the one hand, and the *Cohen's d* effect sizes on the other hand are due to at least two reasons:

1. The way we calculated Cohen's *d*, effect size values do not take multiple testing into account (*df*s are not penalized for multiple testing), while our *p*-value analyses do correct for false

discovery rate (Benjamini-Hochberg's [1995] method). To provide greater clarify, we now also report family-wise-adjusted 95% *CI*s of the effect sizes. Because Benjamini–Hochberg is a sequential method (to our knowledge there is no corresponding adjustment for confidence intervals), we used the multivariate *t* distribution (MVT; via the *mvtnorm* package) to adjust for multiple testing of confidence intervals – a method recommended in biostatistics. The approach we use produces a less conservative adjustment for multiple testing compared to Bonferroni. It uses the same covariance structure as the estimates to determine the adjustment. In the revised Table 1 we can compare inferences from confidence intervals of effect sizes, significance testing and Bayes Factor (analyses with weakly informative priors, as recommended by Andrew Gelman and others). These results are very similar. In only for three domains, 95% CIs of the effect size do notinclude zero. Moreover, in only one domain (i.e., life satisfaction) the lower bound of the 95% *CI* of the effect size is not negligible ($d > .10$), mirroring Bayes Factor results.

2. Averaging *d*s is inappropriate in our context because teams varied in the number of domains they considered and responses were interdependent– some teams submitted forecasts for a few domains, whereas others submitted forecasts for all twelve domains. Consequently, a central tendency of effect sizes across twelve domains would need to take the unbalanced nature of the data and interdependence into account.

In the revision we have added MVT-corrected confidence intervals for effect sizes, and decided to highlight the interpretation of the Bayes Factor by including relevant output in Table 1. Further, we now direct the reader more firmly to the supplementary analyses directly comparing results from different statistical models (Multiverse analyses of domain-general accuracy). We added a new Figure S15 and the description of this most basic form of analysis—inspecting forecasts (lowess curves of scientists and the naïve crowd) against the ground truth trends in the "multiverse analyses" section: for most domains visual inspection of lowess curves from each group suggests a negligible difference. We also continue highlighting domains producing significant effects and showing evidence in support of the difference when inspecting the Bayes Factor, and when examining the Figure S15.

Relatedly, it would help to clarify exactly what analysis is being run on P10 "Comparing accuracy of scientific and lay forecasts, we observed no significant difference between social scientist and lay crowds, F(11, 1747) = 0.88, P = .349, R2 < 0.001" (this may also apply to other reported results, I would clarify the analyses being run)

We have added a clarification about the (full) linear mixed effects model in the results (further information is also included in the supplementary materials discussing each model in greater

detail). Throughout the manuscript, we have added relevant details about analyses for all other results.

Figure 1 plots only the best performing naïve statistical benchmark per domain. Considering the variability in performance between these benchmarks by domain (Figure 2), is this a fairer comparison than plotting the mean or median benchmark performance? I'd make a similar point about the results at the top of P14. Having to beat all three benchmarks seems like a tall order given the seemingly random variability in performance between the benchmarks in any given domain (although perhaps there's an argument that the scientists should be easily beating all three). Overall, these results left me with the impression that the scientists were consistently underperforming the benchmarks, when Fig S2 seems to show scientists being competitive (or better than) the benchmarks quite regularly. I think it might be more informative to either come up with some composite of the benchmarks to compare against, or report performance of scientists vs each benchmark separately.

Thank you. Once again, this is a topic members of the Forecasting Collaborative debated at length. First, we should note that Figure 2 reports tests against individual benchmarks, thus information about differences from each marker are presented in the main manuscript. Further, when we report performance against the top naïve benchmark, subsequent estimates of ratio scores and the corresponding confidence intervals reflect ranges of statistical values for each benchmark comparison. We chose not to discuss and compare performance for each benchmark in more detail to enable parsimony in reporting and avoid false positives (due to multiple testing). As we elaborate below, the picture remains quite similar when examining the averages across three benchmarks, too. To be consistent across our analyses, in each tournament we perform a correction for simultaneous inference of estimates across the 12 domains.

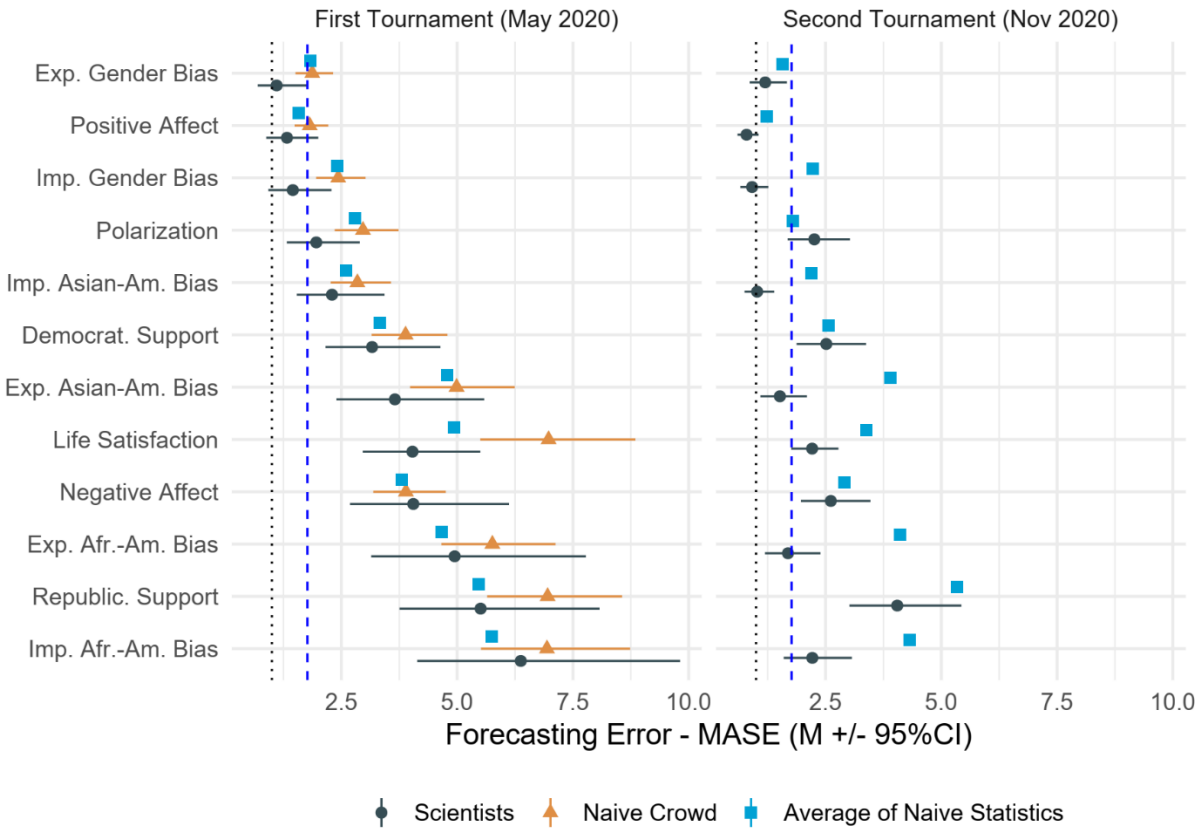
In the process of writing the manuscript, we initially considered information for each of the three benchmarks presented in Figure 1, but it appeared cluttered. Some members of Forecasting Collaborative favored a more parsimonious approach that precludes multiple comparisons (and avoids false positives). Consequently, we moved individual benchmark estimates to the supplementary materials (see Figure S2) and focused on the top benchmark in the main text.

The logic behind our analyses is that scientists should perform better than *any* of the three naïve benchmarks. An analogy would be a set of standards for evaluating quality control in industry. When distinct diagnostic tests in quality control are set up, it is not sufficient for a procedure to pass these tests *on average* (which can happen even if performance is subpar on two out of three tests). Consider three distinct benchmarks testing performance of a nuclear reactor. Given that the tests are distinct in nature and probe against different aspects of overall performance, it

would not be very informative to know whether the reactor passes such tests on average—even if one of the tests is not passed, it may still lead to a nuclear meltdown! Additionally, measures of central tendency performed across a small number of heterogeneous tests can be unreliable.

We selected three benchmarks based on their common application in the forecasting literature (historical mean and random walk are most basic forecasting benchmarks) or the behavioral / social science literature (linear regression is the most basic statistical approach to test inferences in sciences). Moreover, these benchmarks target distinct features of performance (historical mean speaks to the base rate sensitivity, linear regression speaks to sensitivity to the overall trend, whereas random walk captures random fluctuations and sensitivity to dependencies across consecutive time points). Each of these benchmarks may perform better in some but not in other circumstances. Consequently, to test the limits in scientists' performance, we examine if social scientists' performance is better than each of the three benchmarks.

Nevertheless, below is the graph with the average across three benchmarks. As you can see, in the First Tournament, for 9 out of 12 domains scientist forecasts 95% CI included the naïve statistic average. In the Second Tournament, this was the case for 5 domains. These inferences are not that different from what we find if we examine if scientists performed better than all three benchmarks (T1: 10 out of 12; T2: 7 out of 12), and which we report in the main text. The overall message remains similar: Across domains, the average of naïve forecasting methods produced a similar amount of error compared to social scientists' forecasts in Tournament 1. The effects are more nuanced in Tournament 2. In the revised supplement, we added this figure and referred to it when discussing the rationale for our chief analyses, as well.



To address the reviewer's point in the context of our response above we did the following:

- We added a new section to the supplement titled "Rationale for testing performance against three naïve benchmarks", where we elaborate on our logic of testing performance against three benchmarks
- We included a new section in the results discussing differences against the average of the three benchmarks. Here, we estimated means averaged across three benchmarks from the linear mixed effect model, in which forecast/benchmark ratio scores for each of the three benchmarks are nested in participants, thus controlling for interdependence of estimated differences between domains – it's the same approach we used for estimating ratio of performance against benchmarks in our main analyses; we clarify our statistical approach in the main manuscript. We included this section to ensure the reader has complete information and can decide on their own how much merit to put into estimates from the average of naïve MASE scores. We also included the graph with average scores across naïve statistic in the supplement.

- We conclude the revised results section ("How accurate were behavioral and social scientists at forecasting?") by stating: "Overall, social scientists tended to do worse than the average of the three naïve statistical benchmarks in Tournament 1. While scientists did better than the average of naïve benchmarks in Tournament 2, this difference in overall performance was quite small, *M*(forecast /benchmark inaccuracy ratio) = 1.43, 95%*CI* [1.26; 1.62]. Moreover, in most domains at least one of the naïve benchmarks was on par if not more accurate than social scientists' forecasts."
- We have updated the Discussion section, accordingly: "scientists' original forecasts were typically not much better than naïve statistical benchmarks derived from historical averages, linear regressions, or random walks, except for a few domains concerning racial and gender-career bias."

MASE was a new metric to me (and I'm not an expert in any of these naïve forecasting benchmarks), so I appreciated the guidelines for interpreting it. To be clear, is it correct that a MASE < 1 (e.g. in Figure 1) would be quite impressive in the current context? (given that 1.38 is considered "good" prediction in tourism, and 1.76 being better than median data science competitions). I assume this is because the one-step random walk is benefitting from being in-sample, vs the out-of-sample predictions provided for the whole year from scientists? If I'm understanding that correctly, I think it may be worth mentioning something along those lines as a caveat in the results reporting. On my initial pass I got the impression that not hitting MASE < 1.0 would be quite meaningful, but after digging in to the calculation a bit more I'm thinking it's actually quite a high bar in this context.

Thank you for pointing it out. Opinions on MASE < 1 as a benchmark in the forecasting community are mixed - it is reasonable to use when the historical data is expected to inform future trends, following a similar pattern, which may be a questionable assumption with the way historical trends unfolded over the first six months of the pandemic, see new Figure S15). We do agree with the Reviewer that it may be a high bar and so we used additional empirically derived benchmarks from other forecasting tournaments. We have now added a caveat:

"One should note that inferior performance against the in-sample random walk (MASE > 1) may not be too surprising; errors of the in-sample random walk in the denominator concern historical observations that occurred before the pandemic, whereas accuracy of scientific forecasts in the numerator is compared concerns the data for the first pandemic year."

Of course, it's entirely possible that there are good responses to each of these points, but I'd like the authors to consider them and make sure the conclusions are well calibrated to the data.

We have responded to each of these insightful queries and revised the manuscript and the supplementary materials accordingly.

Then, more minor suggestions that some readers may find helpful:
From the text, it's not clear to me what tests/comparisons were pre-registered and which were exploratory. Please clarify this.

In the revision, we have restructured the manuscript (initial submission was directly forwarded from *Nature*, and methods were in the supplementary materials), highlighting the pre-registration section at the beginning of the revised Method section. Our pre-registration included: i) key research questions; ii) data processing; iii) locking-in forecasts of participating teams; iv) use of key metric (MASE) for evaluating performance across domains, v) naive benchmarks (e.g., simple interpolation algorithms); vi) comparison of forecasting approaches; vii) examination of opportunity to update forecasts; and viii) types of covariates we consider in analysis of exogenous variables that may enhance accuracy (e.g., confidence, conditional factors and counterfactuals, number of team members, disciplinary diversity). We also pre-registered data analytic procedures, including how we categorized forecasts in terms of method, categorization of additional parameters in the model, teams, and update justifications (https://osf.io/u9x4m). In addition, we pre-registered comparisons against naïve benchmarks (naïve model in forecasting literature is used synonymously with a random walk; we also included historical mean as another frequently mentioned naïve method). Further, we pre-registered a two-tailed comparison of MASE scores across forecasting types (purely theoretical, purely data-driven and hybrid models) in linear mixed models (MASE scores nested in teams), and a contrast of theory-free models to theory-inclusive models and use of post-hoc pairwise tests for evaluating accuracy.

We did *not* pre-register use of a lay crowd sample prior to collecting their forecasts in June 2020 (but we did pre-register this sample in early September, 2020, prior to cleaning or evaluating their data) and we deviated from the pre-registration in testing all individual predictors (e.g., team characteristics, model simplicity, number of parameters in the data model) simultaneously, instead of performing separate analyses.

We explain the above in the relevant section of the revised methods section, and refer readers to the pre-registration plan we initially submitted to *Nature Human Behavior* for review in May 2020, and which is posted on the Open Science Framework:

"Pre-registration and deviations. Forecasts of all participating teams along with their rationales were pre-registered on Open Science Framework (https://osf.io/u9x4m). Additionally, in an a priori specific document shared with the journal in May 2020, we outlined the operationalization

of the key dependent variable (MASE), operationalization of covariates and benchmarks (i.e., use of naive forecasting methods), along with the key analytic procedures (linear mixed model and contrasts being different forecasting approaches). We did not pre-register the use of a Prolific sample from the general public as an additional benchmark before their forecasting data was collected, though we did pre-register this benchmark in early September, 2020, prior to data pre-processing or analyses. Deviating from the pre-registration, to protect against inflating *p*-values, we performed a single analysis with all covariates in the same model rather than performing separate analyses for each set of covariates. Further, due to scale differences between domains, we chose not to feature analyses concerning absolute percentage errors of each time point in the main paper (but see corresponding analyses on the GitHub site of the project https://github.com/grossmania/Forecasting-Tournament, which replicate the key effects presented in the main manuscript).“

On page 9, it may help to clarify that participants were predicting future values for the same exact polls/data that they were provided as historical data. (this is obvious in the supplemental materials, but wasn't obvious to me from the text)

We now clarify this as follows: “Participating teams received historical data that spanned 39 months (January 2017 to March 2020) for Tournament 1 and data that spanned 45 months for Tournament 2 (January 2017 to September 2020), which they could use to inform their forecasts for the future values of the same time series.”

On P10, the results in the first paragraph (comparing social scientists vs in-sample random walks and prior forecasting competitions) are reported descriptively, whereas the next paragraph (comparing social scientists to lay crowd) rely on inferential statistics. Is this as intended?

Yes, it is intentional, precisely for reasons this Reviewer indicated earlier. We did not want to emphasize the high-bar benchmark (in-sample random walk). Further, prior forecasting competitions are interesting, and the heterogeneous nature of forecasting domains in T3 is informative, but it still concerns a forecasting competition which followed a quite different structure. Several forecasting experts on the team of our Forecasting Collaborative raised the issue with featuring it too prominently and consequently we decided to merely report these markers in a descriptive fashion, as holistic guidelines rather than specific tests.

Potentially useful citation with some counter-points to the Van Bavel et al. paper [47], and echo some of the points in the current discussion:
IJzerman, H., Lewis, N.A., Przybylski, A.K. et al. Use caution when applying behavioural science to policy. Nat Hum Behav 4, 1092–1094 (2020).https://doi.org/10.1038/s41562-020-00990-w

Thank you. We are aware of this paper and have even cited it in other projects of the Forecasting Collaborative. We have included it in the revised Discussion.

the Van Bavel paper, and some others, appear only in the references list and are not cited in the text.

Van Bavel et al. was cited in the Methods section (citation 48):

"Our domain selection was guided by the discussion of broad social consequences associated with these issues at the beginning of the pandemic [48,49], along with general theorizing about psychological and social effects of threats of infectious disease [50,51]"

In addition, we have double-checked and updated the Reference section and updated it where necessary.

Table 1 seems to be missing the 95% CI for the d scores.

We added the 95% CIs for the effect sizes.

Thanks to the authors for this massive effort.

Thank you so much, we appreciate your valuable feedback!

Response to **Reviewer #2:**

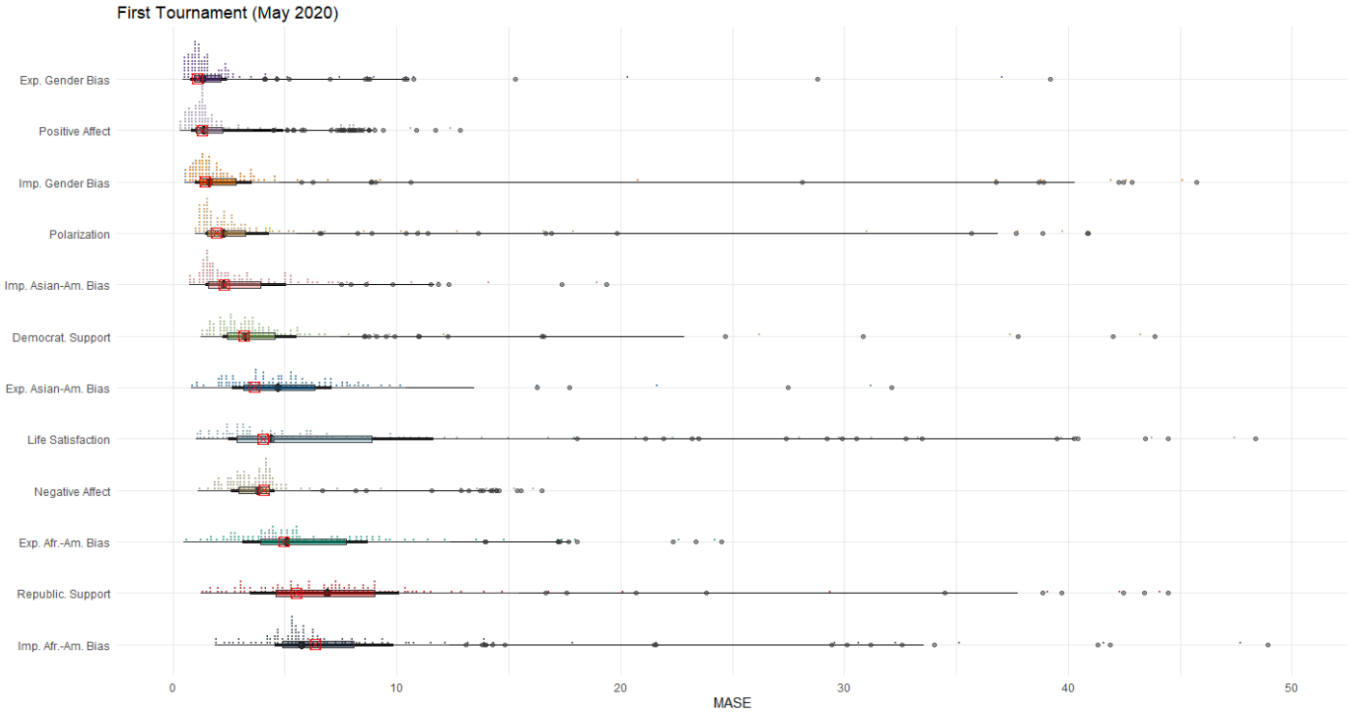
Overall, I want to thank the authors for this interesting paper, and I want to especially thank them for doing a mass collaboration. Team science is challenging for many reasons, but it can lead to really interesting research, such as this paper.

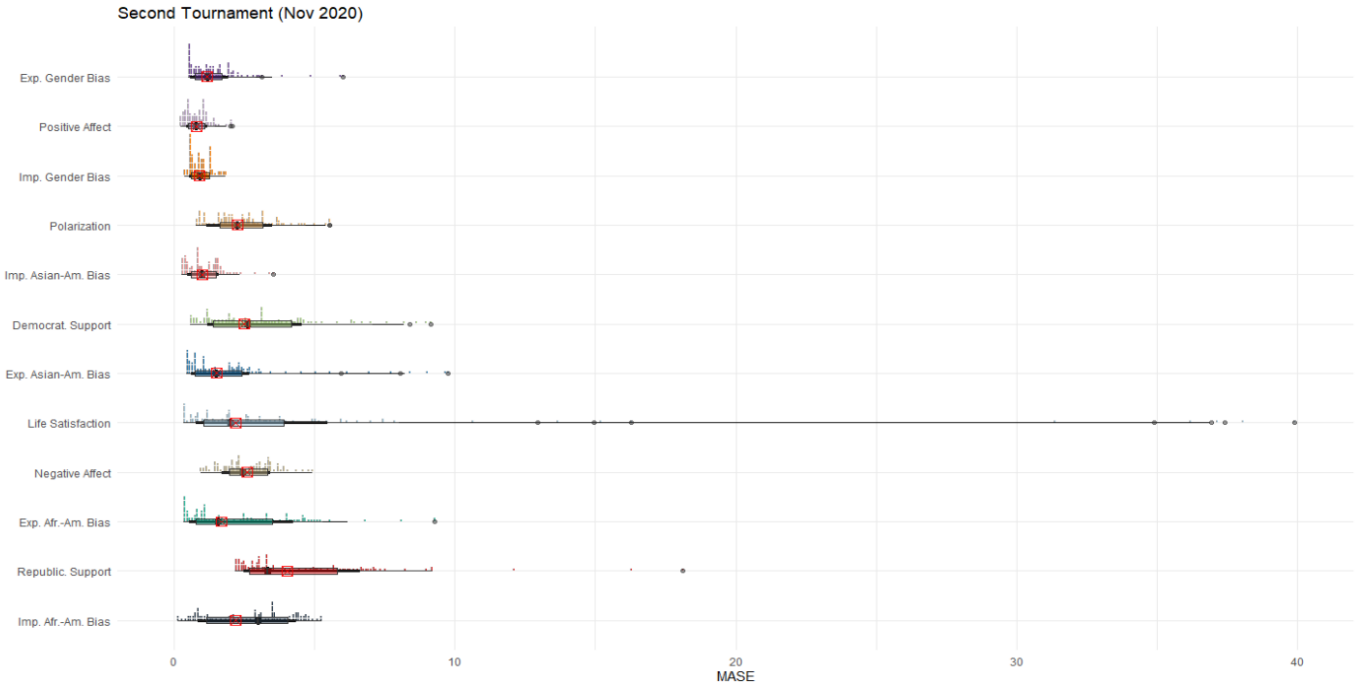
Thank you, both for the support and for raising important issues below! They gave us an opportunity to clarify our methods and results.

I have several questions/concerns/suggestions/provocations that I hope will help the authors improve the paper. These are grouped in major (in the sense that I really think they should be addressed) and minor (in the sense that I think they may be addressed or not, as the authors and editors see fit). They are listed in no order.

Major:
- Heterogeneity within participants. One of the main ways that this differs from other projects using the common task method is that the authors don't simply present results from the best forecaster, they present average results. This means that a small number of bad forecasters could pull down the average performance of the experts. I would like to understand more about the heterogeneity of responses. Are a small group of bad "experts" pulling down the overall score? Perhaps this is represented in Figure 1, but it is not clear to me what the intervals are around the "scientists" predictions. Perhaps something like Fig 3 could include the empirical distribution of errors (not just the standard deviation) or could be presented in terms of median. Naturally, I would leave it to the authors to come up with the way they think is best to address this concern.

This is an important point – as we reported in the initial manuscript (note of Table 1), MASE scores from scientists (and lay people) were indeed skewed to the right. Below are the "box-and-dots-plot" graphs for each tournament depicting the forecasting errors (MASE scores) via dot plots with median and quantile interval, and the overlaid boxplot (median/interquartile ranges and outliers) for each forecasted domain. The scores for some domains are indeed somewhat skewed. To account for this skew, we probed the effectiveness of different transformations (via Q-Q plots of residuals from respective models) and settled on performing all key MASE analyses on log scores. By using logs, analyses become less sensitive to outliers. As reported in the original figure legends, we subsequently back-transform these scores to obtain estimated means per domain, as well as corresponding confidence intervals. The reason we relied on model-based estimates was to account for the interdependence in responses forecasting teams provided for different domains they considered. In short, the presented estimated means and 95% CIs in Figure 1 take the non-normal distribution of initial responses into account when performing relevant comparisons against various benchmarks. In the graphs below, these estimated means are presented with a red crossed square on the graphs below: For most domain, these estimates were either very close to the median forecast or presented an overestimation of social scientists' performance (lower MASE score than suggested by the medians – 3 domains in Tournament 1 / 2 domains in Tournament 2). Only in two cases (implicit African American bias in Tournament 1, Republican support in Tournament 2) median MASE scores were slightly lower than estimated model parameters, but these differences were negligible.


First Tournament (May 2020)
Exp. Gender Bias
Positive Affect
Imp. Gender Bias
Polarization
Imp. Asian-Am. Bias
Democrat. Support
Exp. Asian-Am. Bias
Life Satisfaction
Negative Affect
Exp. Afr.-Am. Bias
Republic. Support
Imp. Afr.-Am. Bias
0
10
20
30
40
50
MASE


Second Tournament (Nov 2020)
Exp. Gender Bias
Positive Affect
Imp. Gender Bias
Polarization
Imp. Asian-Am. Bias
Democrat. Support
Exp. Asian-Am. Bias
Life Satisfaction
Negative Affect
Exp. Afr.-Am. Bias
Republic. Support
Imp. Afr.-Am. Bias
0
10
20
30
40
MASE

We clarify our position by adding the following to the manuscript:

- text to the Figure 1 legend: "Social scientists' average forecasting errors, compared against different benchmarks. Scientists = estimated means and 95% Confidence Intervals from a linear mixed model with domain as a predictor of forecasting error (MASE) scores nested in teams. To correct for right skew, we used log-transformed MASE scores, which are subsequently back-transformed when calculating estimated means. In each tournament, confidence intervals are adjusted for simultaneous inference of 12 domains," as well as in the description of the procedure (clarifying that analyses are performed on *log*(MASE) and log of ratio scores).
- relevant text for naïve benchmark analyses in the main text (paragraphs 2-2 in the results section)
- section "Distribution of Social Scientists' MASE in each Tournament" and Figures S12-S13 to the supplement, showing distribution of responses in each tournament, including medians and estimated means from *linear mixed model* analyses (on log MASE scores).

Timing. It is not clear to me exactly which months are being forecast. The start dates are shown on page 9 lines 276-277. This should be clearer. Also, for future readers and readers from outside the United States, I think it should be clearer that this was perhaps an unusual time in the US due to COVID and the Black Lives Matter protests. Perhaps some kind of timeline figure could be added to the SM, which would also include key historical events happening during these periods (e.g., George Floyd murder, COVID vaccine rollout, etc)?

We have added the exact months that participating teams were asked to forecast on the second page of the introduction, as well. Further, we added a timeline of historical events and pandemic impact (COVID deaths) in the supplementary materials (see Figure S14) and made a reference to it in the main text.

Also, I think the authors have the capacity to check empirically whether their outcomes moved in unusual ways during this time.

To address reviewer's point, we examined if societal change trends in each domain were unusual during the time immediately following the first pandemic lockdowns in the US. To this end, we computed the mean absolute difference between observations for the May-Oct 2020 6-month period (the first six months of the Tournament 1) and the average of the last 3 time points of the historical data (Jan-March 2020). We then repeated this procedure for the next 6 months period (Nov 2020-Apr 2021), again computing the difference from the average of the last 3 time points of the historical data. The results showed that the absolute difference was significantly higher across domains for the initial 6-month period than for the subsequent

6-month period, paired *t*-test across domains for the two time points ($df = 11$) = 2.474, $P = .03$. We report these results in the new supplementary section "Societal change over the pandemic."

Because we are not entirely sure what the reviewer means by "unusual ways" and which outcomes they refer to, we also want to briefly touch on additional pragmatic constraints. First, if the reviewer asks about the contribution of historical events to forecasting accuracy, we should note that our key measure of forecasting accuracy (MASE) is computed *across* multiple months, which makes it challenging to examine whether forecasts would differ as a function of specific events that happened within these months. Second, if the reviewer asks us to consider the impact of specific events on the time series of each of the 12 domains, an additional challenge is the subjectivity in determining event duration. For example, consider the death of Ruthor Ruth Bader Ginsburg on Sept 18, 2020. Though we can put an event marker for the month of September 2020 and compare months before and after this date, this may not be the best way to determine the impact of RBG's death on the US society. Arguably, the consequences of her death became most salient in the summer of 2022, with the new US Supreme Court abortion ruling. Conversely, consider singular events like Biden taking office on Jan 20, 2021. Would we consider this event for the duration of the month of January? If so, how would we separate this event from the other major issues happening in US politics in the same months (e.g., Capital Insurrection on Jan 6 or Second impeachment of Donald Trump shortly thereafter)? Given that forecasts were done for each month (rather than for each day), consideration of short-term events happening in the same month would make it impossible to differentiate between distinct historical issues happening in the same month.

Benchmarks. I thank the authors for thinking carefully about benchmarks. This is important and unfortunately relatively rare. When constructing the historical mean benchmark, I don't understand why the authors didn't use the historical mean (they used random resampling instead).

We thought extensively about the appropriate ways to calculate these benchmarks. As we described in the methods section, "estimates from this approach are equivalent to the historical mean for each domain, but also give a sense of the expected range around that mean." We agree that an alternative way would be to simply calculate the historical average, and then estimate the MASE scores from this historical prediction. This type of benchmark cannot capture the variability that one might expect due to chance from analyzing many teams making predictions in different ways, but it has the advantage of being extremely simple to calculate, with no additional assumptions about the distribution of the data. Notably, the resampling approach that we take produces almost identical mean predictions for each of the domains (see Table S8 in the revised supplement; comparison for 12-month Tournament 1 is copied below). Further, for most

domains, the MASE scores from the resampling mean and simple historical mean approaches were close to identical. We do note two domains where this pattern of MASE scores slightly diverged (i.e., explicit gender bias and life satisfaction). Here, the resampling approach produces noticeably less accurate forecasts (i.e., higher MASE scores) compared to the mean of the historical data, despite producing nearly identical mean predictions. Thus, similar performance of scientists to the resampled historical mean for explicit gender bias poorer performance compared to resampled mean benchmark for life satisfaction (see Figure 2 and Figure S2) can be considered underestimations. For these domains, scientists on average did worse than one would have done by taking a historical average of the last 12 data points. In short, the main take home message remains close-to-identical.

We have done extensive testing to identify the cause of discrepancy between MASE (but not forecasting) estimates, and although we have not identified a clear cause for these two discrepancies, we speculate that it may arise from idiosyncrasies in the distribution of the historical data compared to the forecast data. Given the close correspondence between the historical mean benchmark and our simulation approach, as well as the advantage from our simulations of computing MASE based on distributions of expected outcomes, we have opted to retain the current approach (this also mirrors the other two naïve statistical approaches that also rely on simulations). However, if the editor feels that it would add to the results to include the simple historical mean, we would be happy to do so. For now, we added a new section "Comparison of two naïve benchmarks: Resampled historical mean versus averaging of historical data" to the supplementary materials, in which we present similarities and differences of the two approaches and alert readers to the idea that our estimate may in fact be an underestimation of the benchmark performance (and overestimation of scientists' performance compared to this benchmark).

Domain	Historical Mean	Resample Mean	|Diff|	*log*(MASE) historical mean	*log*(MASE) resample mean	|Diff|
Explicit Bias – Af Am	-0.035	-0.035	0.000	1.718	1.713	0.005
Explicit Bias – Asian	-0.037	-0.037	0.000	0.932	0.925	0.006
Explicit Bias – Gender	0.837	0.837	0.000	0.184	0.381	0.197
Implicit Bias – Af Am	0.319	0.319	0.000	1.726	1.725	0.001
Implicit Bias – Asian	0.386	0.386	0.000	0.404	0.410	0.006

Implicit Bias – Gender	0.365	0.365	0.000	0.241	0.326	0.085
Support for Democrats	43.291	43.299	0.008	1.348	1.414	0.066
Support for Republicans	36.721	36.708	0.013	1.845	1.855	0.010
Life Satisfaction	6.362	6.362	0.000	0.304	0.998	0.694
Negative Affect	0.973	0.974	0.001	1.453	1.455	0.002
Polarization	79.385	79.391	0.005	1.021	1.091	0.070
Positive Affect	-0.973	-0.973	0.000	0.310	0.333	0.023

Also, in Figure 1 and Figure 4, I don't understand why the M4 tournament performance is a reasonable benchmark. To be clear, I'm not saying that it is not, but I'm not familiar with the M4 tournament, and the little I do know makes it seem quite different.

Our rationale to use estimates from M4 as a benchmark was two-folds: i) forecasted domains in this tournament were highly heterogeneous; thus, we can assume at least some generalizability to social issues in our forecasting tournament; and ii) it is the most established, once-in-a-decade, tournament in the forecasting community, which includes MASE markers of each team (hence we could use them to calculate median MASE scores of performance in the M4 tournament).

As we described in the Supplementary Method (last paragraph of the "Mean Absolute Scaled Errors (MASE) as a Marker of Forecasting Accuracy" section): "Results of a more heterogeneous M4 forecasting competition of 100,000 real-life time series suggest a threshold 1.89 for the out-of-sample naïve (random walk) statistical benchmark, a threshold of 1.70 for the Theta statistical benchmark, and a threshold of 1.77 when considering the median of all M4 forecasting teams."

M4 relied on a vast number of time series forecasts and a wide range of domains that made up the tournament. More specifically, as Makridais and colleagues wrote in the description of the M4: "The 100,000 time series used in the M4 were selected from a database (called ForeDeCk) compiled at the National Technical University of Athens (NTUA) that contains 900,000 continuous time series, built from multiple, diverse and publicly accessible sources. ForeDeCk emphasizes business forecasting applications, including series from relevant domains such as industries, services, tourism, imports & exports, demographics, education, labor & wage, government, households, bonds, stocks, insurances, loans, real estate, transportation, and natural resources & environment (Spiliotis, Kouloumos, Assimakopoulos & Makridakis, 2020)."

In the supplementary materials, we have added further justification for using M4 estimates as a possible benchmark. We also note that we deliberately avoided making formal comparison to M4 beyond outlining this benchmark line in the Figures and descriptive mentioning of it in the results section, because the context of tournament matters and our context is somewhat different from M4. If the editor prefers, we would be happy to remove this benchmark; we chiefly added it because several members of the Forecasting Collaborative community raised the question of possible benchmarks from prior tournaments and to our knowledge this is the best possible (albeit imperfect) comparison to date.

Construct validity of outcomes: Some of the outcomes are complex measurements on non-random samples. For example, affective well-being is a complex summary of Twitter. The authors seem aware of these issues, but I wonder how much of the forecasting challenges are created by measurement challenges.

We agree and have discussed this point in footnote 23 of the original submission, which we now moved to the main text: "Differences in forecast accuracy across domains did not correspond to differences in quality of ground truth markers: Based on the sampling frequency and representativeness of the data, most reliable ground truth markers concerned societal change in political ideology, obtained via an aggregate of multiple nationally representative surveys by reputable pollsters, yet this domain was among most difficult to forecast. In contrast, some of the least representative markers concerned racial and gender-bias, which came from Project Implicit—a volunteer platform that is subject to self-selection bias—yet these domains were among the easiest to forecast. In a similar vein, both life satisfaction and positive affect on social media were estimated via texts on Twitter, even though forecasting errors between these domains varied. Though measurement imprecision undoubtedly presents a challenge for forecasting, it is unlikely to account for between-domain variability in forecasting errors (Figure 3)."

We also want to stress that all teams were provided information about the sources for the historical and ground truth markers on each excel sheet they used for submissions (see example in the supplementary Appendix).

I am very surprised that accuracy was better for predictions further the future (line 413). I also didn't understand the proposed explanation in the text and Fig S1. Do the authors think this will be a general pattern in other forecasting tournaments or might it be specific to this time period, which the authors acknowledge is unusual?

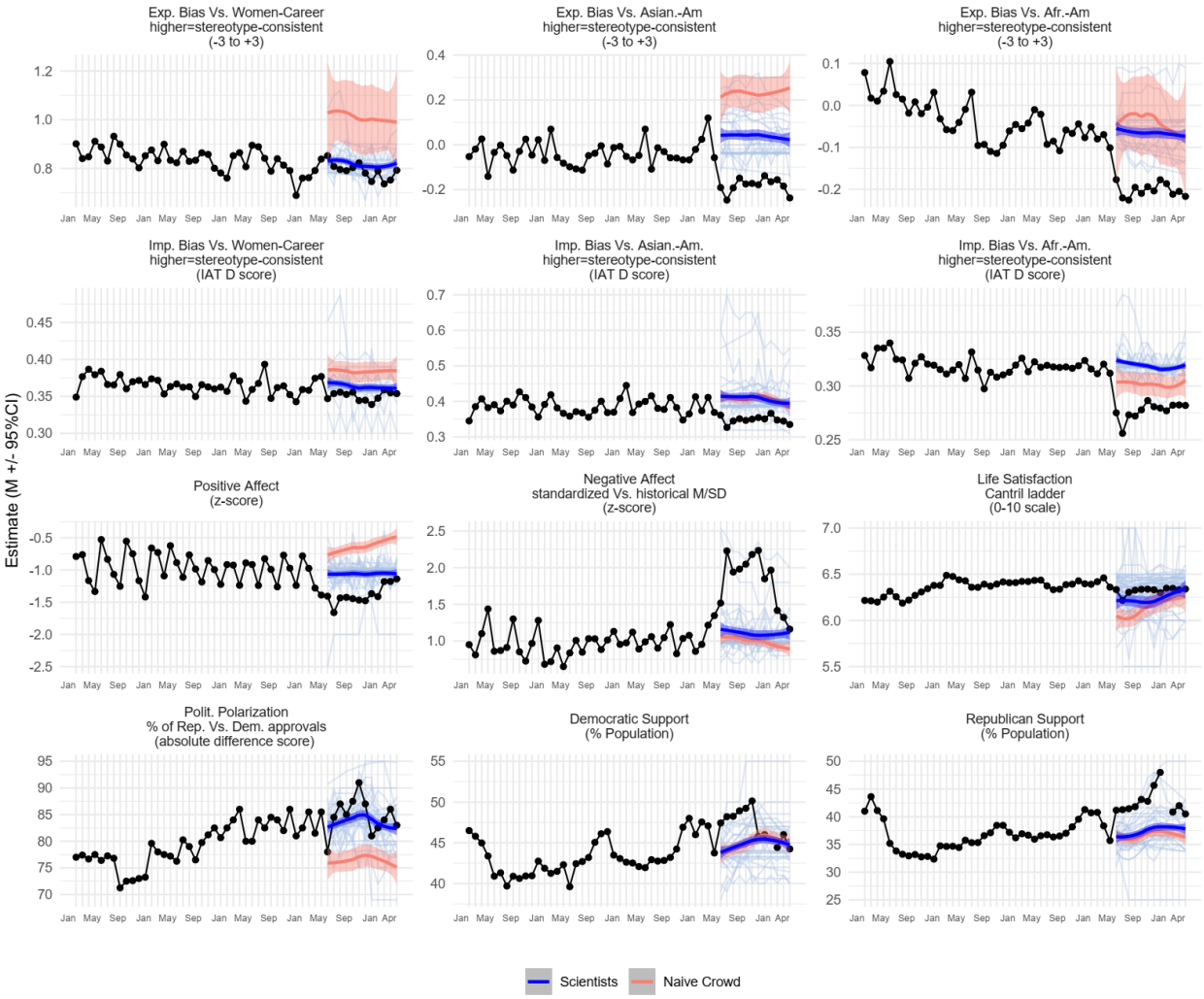
To help clarify this point further, in the revision we plotted predictions (individual and lowess curve across them) for each domain, along with historical data for the last 39 months before the

tournament and the 12 months after May 2020 (see the graph below). It is clear that in many domains (i.e., gender-career bias, positive affect, negative affect, explicit bias against African Americans and Asian Americans, and political polarization), scientists' average forecasts anchor closer to the average of the last three historical data points they were provided (January-March; note that scientists did not receive the April point, as it was not available at the time of the tournament launch). For other domains this pattern is more complex. The pattern of results goes against the hypothesis that the reason people did better in tournament 2 (6 months) instead of tournament 1 (12 months) is because the 12 months tournament included forecasts for time points further temporarily removed. We report this graph in the revised supplement, and we would be happy to move it to the main text if the editor thinks it to be helpful to sharpen the main message of the manuscript.

We agree with the reviewer that this result may on the surface be counterintuitive to the idea that the accuracy will go down with temporal distance and that is why we introduced this result by starting with "Contrary to this seemingly straightforward hypothesis." But in hindsight, our surprise may be somewhat misplaced. For instance, research on focalism bias in affective forecasting and temporal distance research in social judgments would suggest that forecasts for more immediate time points may be more sensitive to local features such as most recent historical data points, whereas forecasts for further removed time points may be more likely to rely on the bigger picture (base rates). Alternatively, as the reviewer suggests, it may be that there is something special about the time period itself. If this is the case, it may be that values within a domain deviated more substantially early on in the pandemic and then returned to baseline. To the extent that predictions by our forecasting teams were anchored on that baseline, they would be expected to show less error later compared to earlier. To test this latter hypothesis, we characterized the average absolute difference between the first 6 months of observations and the average value of the last 3 months of historical data, as well as the absolute difference between the last 6 months of observations and the last 3 months of historical data, across each of the 12 domains. As expected if there is something anomalous about the initial 6 months of the pandemic, we found that on average the divergence was larger in the first 6 months than the second 6 months, paired t across 12 domains = 2.474, P = .03. In other words, on average, we find that across domains, observations may have shifted more substantially during the initial COVID-19 waves and then returned to baseline, explaining why our forecasting teams' initial 12-month predictions, which largely show a trend toward centering on the baseline, became more accurate at more temporally removed points.

Therefore, in the revised results we added these insights (which we report in full in the supplement): "Supplementary analyses further show that many domains showed unusual shifts (vis-à-vis prior historical data) in the first six months of the pandemic, and started to return to the

historical baseline in the following six months (Fig. S15). For these domains, forecasts anchored on the most recent historical data were more inaccurate for the May-October 2020 forecasts compared to the November 2020-April 2021 forecasts."



Also, do the authors have other evidence to support their anchoring hypothesis from the crowd or from the month-by-month predictions?

Yes, the graph above (which is in the Supplementary materials) shows similar (if not more pronounced) anchoring among naïve crowds in Tournament 1; the effect is evident in nine domains. In the supplement, we also point out this effect among lay samples. Note, anchoring

itself cannot fully explain why performance in Tournament 2 is superior to Tournament 1, and so we explore this issue further in the Results section entitled "Comparison of accuracy across tournaments".

Relatedly, in the next paragraph, the authors seem to conclude that forecast accuracy is a function of the amount of timeseries data available. I find it surprising that six months of additional data would make much difference, given that they already had 39 months of data (if I understand the design correctly). If this is true, it suggests that the length of the historical data window is an important design consideration, and then I would wonder why the authors picked the historical data window they did.

We did not mean to convey that "*forecast accuracy is a function of the amount of timeseries data available.*" To clarify our position, we have now added: "we tested whether providing teams additional six months of historical trend capturing the onset of the novel pandemic at Tournament 2 may have contributed to lower error compared to Tournament 1. To this end, we compared the inaccuracy of forecasts for the six months period of Nov 2020-April 2021 done in May 2020 (Tournament 1) and when provided with more data in October 2020 (Tournament 2). We focused only on participants who completed both tournaments to keep the number of participating teams and team characteristics constant."

As we discussed in the response to the previous query, the historical trends diverged more from the historical baseline in the first six months compared to the latter six months. By obtaining feedback about such anomalous patterns, teams in Tournament 2 (Nov 2020) were able to make more accurate forecasts compared to Tournament 1. Recall, all teams received feedback on their prior performance and had an opportunity to revise their initial forecasts.

Discussion: I wonder if the authors have any new theories about which domains are more difficult to forecast based on these results. Based on around line 527 the authors seem to think that this is mainly related to historical volatility. Could this be tested directly?

In the revised manuscript we report this test in the second paragraph of the "Which domains were harder to predict?" result section:

"Domain differences in forecasting accuracy corresponded to differences in the complexity of historical data: domains ranked more variable in terms of standard deviation (*SD*) and mean absolute difference (*MAD*) of historical data tended to have more forecasting error (as measured by the rank-order correlation between median inaccuracy scores across teams and variability scores for the same domain), Tournament 1: $\rho(SD) = .19$, $\rho(MAD) = .20$; Tournament 2: $\rho(SD) =$

.48, $\rho(MAD)$ = .36, and domain changes in variability of historical data across tournaments corresponded to changes in accuracy, $\rho(SD)$ = .27, $\rho(MAD)$ = .28."

We should also note that volatility is not a singular hypothesis, and it can be approached both on the level of cross-domain comparisons and within-domain changes over time (reflecting different levels of analysis), speaking to different types of data complexity. For instance, as we report in the Discussion section, positive and negative affect likely vary in volatility due to psychological differences in complexity of positive versus negative affective experiences – negative experiences tend to be experienced and reported in a more differentiated fashion compared to positive experiences. This has implications for volatility over time, possibly explaining less regular fluctuations in negative compared to positive sentiment on social media in the historical data (see the graph above). To this end, in the Discussion we wrote:

"Domain differences in forecasting accuracy corresponded to historical volatility in the time series. Differences in the complexity of positive and negative affect are well-documented [28,29]."

Further, when discussing differences in forecasting accuracy, we bring up a point concerning volatility in trends during (versus before) the tournament: "Moreover, racial attitudes showed more change than attitudes regarding gender during this period (perhaps due to movements like Black Lives Matter)."

This point is not about between-group differences in historical volatility, but rather about within-domain *change* in attitudes as a function of unique external events.

Besides discussing these different hypotheses, we did not think of compelling theoretical frameworks that immediately accounted for the patterns in the Figure 3. Thus, we refrained from further speculation.

Also, if this is only about volatility that would be important to know so that people don't try to develop theories based on the substantive characteristics of these outcomes.

As we outlined above, it is not solely about volatility/variability in historical trends. Consequently, in the revised discussion we write:

"Moreover, social scientists may benefit from testing whether a given societal trend is deterministic and hence can benefit from theory-driven components, or if it unfolds in a purely stochastic fashion. To achieve this goal, training in recognizing and modeling properties of time

series and dynamical systems may need to become more firmly integrated into graduate curricula in the field. “

The point is that highly volatile time series may not necessarily be deterministic—something that one can evaluate by carefully examining historical time series by decomposing it into trends, autoregressive components, and examining seasonal features to achieve stationarity.

I really like the idea of Fig S1. In some outcomes the ground truth appears to be a natural continuation of the historical timeseries. But for other outcomes like “Exp Bias vs Asian Americans” and “Exp Bias vs African Americans” there appears to be a kind of discontinuous jump. Perhaps I’m reading too much into just two points in the historical data. Do the authors have any ideas about whether there might be sharp changes in some outcomes right around the tournament time and if so why this might be the case?

As we discussed in response to earlier comments, the onset of the pandemic itself appears to have resulted in unusual trends (vis-à-vis) historical baseline. The pandemic is the most likely explanation for the dramatic jump in negative affect on social media, with a slow return to historical baseline by the end of the tournament. We address this issue in the new supplementary section “societal change over the pandemic” and depict historical trends in the new Figure S15.

Additionally, trends on bias-related scores are likely driven by the Black Lives Matter movement (BLM). Note that the historical data over the prior 40 months suggests a downward trend in explicit bias over the last 3 years before the pandemic, so the drop in explicit bias against Asian and African Americans is in some ways a continuation of the trend, likely accentuated by the BLM protests in the summer (as we suggest in the Results section).

Minor:
I believe there is a typo on line 1165 (could be 5,000 – 6,000 or 50,000 to 60,000).

Thank you! We have fixed it: it is 50,000 – 60,000.

- The term “expert” is used quite loosely here. Do we think that graduate students are “experts” in social sciences? To be clear, I think it is great to assess the accuracy of graduate students, I’m just expressing concern about the construct validity of the term “experts” given the pool of participants.

We agree that defining expertise in social sciences is complicated. If expertise concerns broad knowledge of social sciences–i.e., the scientific method used in social sciences, general

paradigms in social sciences, etc. (rather than social standing in a discipline), PhD students and their European equivalents have substantially more knowledge about social sciences than lay people. That is why we describe this comparison as the one examining "domain-general expertise," with the term "domain-general" aiming to capture general sense of knowledge about social sciences relative to lay audience. In subsequent tests, we focus on other operationalizations of expertise, including number of members on a forecasting team with a PhD (or its country-specific equivalent), number of publications in the forecasted domain, and subjective confidence of team members in the domain-specific expertise (see Figure 5). In the Discussion section we invoke the idea of subject matter expertise rather than domain-general expertise.

In the revised Methods section, we devote a paragraph to this issue:

"**Team Expertise**. Because expertise can mean many things [2,63], we used a telescopic approach and operationalized expertise in four ways of varying granularity. First, we examined broad, domain-general expertise in social sciences by comparing social scientists' forecasts to forecasts provided by the general public without the same training in social science theory and methods. Second, we operationalized the prevalence of graduate training on a team as a more specific marker of domain-general expertise in social sciences. To this end, we asked each participating team to report how many team members have a doctorate degree in social sciences and calculated the percentage of doctorates on a team. Moving to domain-specific expertise, we instructed participating teams to report if any of their members had previously researched or published on the topic of their forecasted variable, operationalizing domain-specific expertise through this measure. Finally, moving to the most subjective level, we asked each participating team to report their subjective confidence in teams' expertise in a given domain (see Supplementary Information)."

We further state in the revised Discussion section:

These findings, along with a lack of domain-general effect of social science expertise on performance compared to the general public, invite consideration of whether what usually counts as expertise in social sciences translates into greater ability to predict future real-world trends."

By adding this statement, we hope to stimulate a discussion about the definition and meaning of expertise in social sciences and whether forecasting real world phenomena should be considered part of such expertise.

The design that these researchers used was unusual, in part because there are not yet clear standards in this area. I wonder if the authors should reflect on what they learned about designing forecasting tournaments and include that in the article. For example, it seems like participants were able to choose what domains they forecasted, and this probably made the analysis more complex. Do the authors think this complexity was worth it? Also, the authors had two nested tournaments. Is that something they think must be done in the future as well? How would they recommend that future researchers select domains?

Thank you for giving us an opportunity to reflect on our experiences. We have added a paragraph in the Discussion section.

"The nature of our forecasting tournaments allowed social scientists to self-select any of the twelve forecasting domains, inspect three years of historical trends for each domain, and to update their predictions based on feedback on their initial performance in the first tournament. These features emulated typical forecasting platforms (e.g., metaculus.com). We argue that this approach enhances our ability to draw externally valid and generalizable inferences from a forecasting tournament. However, this approach also resulted in a complex, unbalanced design. Scholars interested in isolating psychological mechanisms fostering superior forecasts may benefit from a simpler design whereby all forecasting teams make forecasts for all requested domains.

Another issue in designing forecasting tournaments involves determination of domains one may want participants to forecast. In designing the present tournaments, we provided participants with at least three years of monthly historical data for each forecasting domains. An advantage of making the same historical data available for all forecasters is that it establishes a "common task framework" [9,16,17], ensuring main sources of information about the forecasting domains remain identical across all participants. However, this approach restricts types of social issues participants can forecast. A simpler design without inclusion of historical data would have had an advantage of a greater flexibility in selecting forecasting domains.."

We also note that self-selection of domains is unlikely to be problematic, because our lay participants (benchmark in tournament 1) were instructed to provide forecasts for the same set of 3 domains (randomly assigned to a cluster of thematically related domains) and their forecasts were very similar to those provided by scientists.

I was not able to find the pre-registered analytic plan on Open Science Framework (OSF). I'm sure that this represents an error on my part; I am not familiar with this website. However, I would hope that it can be reviewed. Perhaps in the future the authors could provide a direct link

to the document with the historical timestamp for reviewers and readers who are not as familiar with OSF.

The pre-registered plan was previously submitted to NHB (NATHUMBEHAV-200410305PI) as an inquiry in April 2020; you can find this submitted plan here https://osf.io/7ekfm. Note, it is not a conventional pre-registration. We initially aimed to present the manuscript as a pre-registered report which we submitted to the journal. We subsequently officially pre-registered our methods on September 11, 2020 https://osf.io/u9x4m. The exact copy of this report is uploaded on OSF, but the time stamp suggests a later date (in the chaos of the first COVID lockdown, we failed to upload it to the OSF and subsequently shifted focus to the actual tournament, with reports on GitHub). We have now harmonized the GitHub and OSF parts of the project, so all aspects of the project can be found together. We also added the direct link to the pre-registered plan document, and moved the section about deviations from the pre-registration to the front of the revised methods section.

The authors write "Fig 1 shows that in Tournament 1, social scientists' forecasts were, on average, inferior to in-sample random walks in all domains." (line 322). However, it seems that for some domains the marker for "scientists" is to the left of the market for "naïve statistic". I hope that the match between the text and figure could be clarified.

The naïve statistic marker (the blue square) concerns the out-of-sample random walk, whereas the vertical dotted line refers to in-sample random walk. As we discussed in response to Reviewer 1 (and in the revised text), in-sample random walk is often harder to outperform. We also revised the text, which now states "social scientists' forecasts were, on average, inferior to in-sample random walks in nine domains." This is because correcting confidence intervals for multiple testing resulted in the overlap of the confidence interval of scientists' forecasts and the in-sample random walk benchmark (MASE = 1) in three domains.

Several places around line 323 the authors talk about a "in-sample random walk", and I'm not sure what that is or how that compares the rank walk described on line 315.

We have clarified this point in response to a similar query from Reviewer 1, and updated the manuscript accordingly.

I found it very interesting that there was little overlap between the winners in tournament 1 and tournament 2. However, focusing just on the winners discards a lot of information. Is it the case that teams that did well in tournament 1 did well in tournament 2? If there is a lot of reversion to

the mean that would suggest that luck partially explains some of the good performance in tournament 1.

This question concerns general consistency in responses. In fact, MASE performance across the tournaments showed a significant degree of consistency. To test the extent to which MASE scores in Tournament 1 predicted MASE scores in Tournament 2, one can fit a linear mixed model with T1 MASE scores, domain, and T1 MASE x domain interaction predicting T2 MASE scores among participants who completed both tournaments. As in all other analyses, we convert MASE to logs to correct for non-normal distribution. Further, we scale both predictors and the criterion variable to obtain standardized estimates. Next, we can examine simple slopes by domain. The results below show estimates by domain, along with 95% CIs, adjusted for multiple tests (as in our other analyses). For nine domains, we see a significant positive association of forecasting accuracy across the tournaments, with the effect size β ranging from .16 for negative affect to .33 for explicit gender-career bias.

domain	MASE1_w1.trend	SE	df	lower.CL	upper.CL
eafric	0.3107	0.0611	149	0.1335	0.488
easian	0.2590	0.0518	154	0.1090	0.409
egend	0.3266	0.0497	153	0.1826	0.471
iafric	-0.0502	0.1691	152	-0.5404	0.440
iasian	0.2450	0.0745	154	0.0291	0.461
ideoldem	0.2751	0.0462	155	0.1412	0.409
ideolrep	0.2429	0.0319	155	0.1505	0.335
igend	0.2278	0.0822	146	-0.0106	0.466
lifesat	0.2570	0.0125	153	0.2206	0.293
negaffect	0.1598	0.0968	154	-0.1208	0.440
polar	0.2708	0.0476	154	0.1329	0.409
posaffect	0.3223	0.0895	155	0.0627	0.582

We have added relevant information in the manuscript:

“ Further, results of a linear mixed model with MASE scores in Tournament 1, domain, and their interaction predicting MASE in Tournament 2 showed that for eleven out of twelve domains

accuracy in Tournament 1 was associated with greater accuracy in Tournament 2, *Md*(standardized β) = .26."

Similarly, we observed a high degree of consistency *within* tournaments. We report on consistency within each tournament in the "Consistency in Forecasting" section, showing that "model accuracy in one subset of predictions (ranking of model performance across odd months) was highly correlated with model accuracy in the other subset (ranking of model performance across even months)."

On line 332, I don't understand what this R2 value of less than 0.001 represents.

It is the part R^2 for the scientists vs. lay people contrast main effect from the linear mixed effect model where scientists vs. lay people and domain are main effects, as well as their interaction. It reflects part of the variance explained by the main effect (contrast of scientists and lay people). We have revised the analytical segment of the results to more firmly introduce the statistical method and have also added clarification that this statistic represents a part R2.

I didn't understand the claims on line 336 about Bayesian analysis, and it was not obvious to me where to find the necessary details in the SM.

We have revised Table 1 to provide conventions for interpreting the Bayes Factor scores. Further, the supplementary materials now include a section "multiverse analyses" which describes the procedure and type of priors we used:

"We computed Bayes Factor approximations using *rstanarm* [6] package in *R* [7]. Following guidelines [8], we relied on weakly informative priors for our linear mixed model and used *emmeans* [9] and *bayestestR* [10] packages to obtain Bayes Factors for individual marginal means of social scientists and lay people. Only a single domain – life satisfaction – showed strong evidence in support of greater accuracy of scientists' predictions compared to lay predictions, $BF = 22.72$ (see Table 1 in main text) and eight domains in support of the null hypothesis, where there was moderate to strong evidence that there were no differences between the predictive accuracy of scientist and lay people, $BF \leq 0.12$. For the remaining three domains, there was either weak evidence in support of the difference (explicit and implicit gender career bias), or too little evidence of support for either hypothesis (political polarization). "

Around line 430 I didn't understand what is meant by model accuracy. Is model being used with interchangeably with team?

Correct. It is simply the accuracy of the teams' forecasts. In the revision we have deleted the term model to avoid further confusion.

On line 597 the authors suggest that social scientists could benefit from testing whether a trend is stochastic or deterministic. I agree that this would be very valuable knowledge, but it seems impossible to test (at least to me). If the authors know how to do this, it would be great if they could share a bit more. If this is impossible, then perhaps it should be removed?

We agree that it is not always easy to do, but at the minimum, one can decompose a time series into the trend, autoregressive, and seasonal components, examining each of them and their meaning for one's theory and model. Further, conventional techniques in the time series literature include unit root tests—e.g., the (Augmented) Dickey-Fuller test or the KPSS to differentiate stochastic from deterministic time series. We added a new sentence to this end:

"For instance, one can start by decomposing a time series into the trend, autoregressive, and seasonal components, examining each of them and their meaning for one's theory and model. One can further perform a unit root test to examine whether the time series is non-stationary. "

On line 606 the authors write that the ability to accurately predict trends in these variables "would appear to be of critical importance." It is not clear to me that predicting these trends slightly more accurately than a naïve model is really important. Did the authors include "appear" here as a hedge? Could they explain more about why slightly better predictions would be "of critical importance". Also, just to be clear, I'm not saying that the domains they studied are not important, just that slightly better prediction of those domains is not obviously important.

We did not mean to communicate that it is important to predict these societal issues "slightly better than naïve models." Rather, our concluding paragraph of the discussion aims to return the reader to the bigger picture point – it is generally important to test predictive power of existing theories of phenomena such as prejudice, political polarization or well-being, and social scientists do not do a good job at predicting these phenomena – i.e., they "have a lot of room for growth." The reason we used tentative language "would appear to" is because some scholars in social sciences we have encountered objected that prediction is in fact important to science. We disagree, and use the tentative language to ensure room for a future debate on this issue.

- I really like the idea of Fig S1, but I find it hard to interpret. In the negative affect panel, for example, why does it seem that there are vertical blue bands starting in November?

Thank you for pointing this out. *ggplot2* had produced some mapping errors (without warnings) for multi-factor multi-layered plots, resulting in the artifact - vertically looking lines. We have now corrected this and there should not be any vertical lines. We also added more explanation of the lines in the legend.

I wanted to thank the authors again for an interesting, important, and stimulating article. I think it has the potential to be read by a wide range of scholars, and I hope that the feedback provided above helps make the manuscript clearer and ultimately more impactful.

Thank you very much for your constructive feedback.

Response to **Reviewer #3:**

The purpose of this ms is captured in the following statement:

Line 236-244: "Prior forecasting initiatives have not fully addressed this question [predicting trends in social phenomena] for two primary reasons. First, forecasting initiatives with subject matter experts have focused on examining the probability of occurrence for specific one-time events (4, 6) rather than the accuracy of ex-ante predictions of societal change over multiple units of time (7). The likelihood of a prediction regarding a one-off event being accurate is higher than that of a prediction regarding societal change across an extended time period. Second, forecasting efforts have concentrated on predicting geopolitical (4) or economic events (8) rather than broader societal phenomena. "

Thank you for acknowledging our contribution to the wider forecasting literature.

An enormous amount of material is described. A proper review must check the integrity of the data, the analytical methods and the conclusions.

Indeed, we encourage other scholars to make use of our data for their research. As we outlined in the supplementary methods and in the CrediT Author contribution section, several team members independently verified all data, analytical methods and conclusions ("Validation: K.S., X.E.G., and L.W."). Each analytical step is annotated in the *R* Markdown file. Each file is described in the Readme doc on GitHub, outlined in the original submission https://github.com/grossmania/Forecasting-Tournament . To ensure even greater transparency,

we have now added a section describing the steps of the validation in the Supplementary materials:

"**Code review process**. Members of the Forecasting Collaborative performed an extensive code review during the preparation of the manuscript to assess the reproducibility of the code and cross-validation. Two members volunteered for the code review. The two code reviewers downloaded and ran the data and code in the "R Markdown file of the main analyses" shared through the GitHub repository of the project. The code reviewers carried out the code check on their own computers (i.e., using different operating system environments from where the code was initially developed), and let the main authors know that the code was possible to run, and the code did carry out the intended analyses. The shared code was divided into sections that correspond to the statistical output and visualization in the manuscript. If the code reviewers spotted an error, they were instructed to take a note of the line and then write a short description that explains what the potential issues are. Any typos and coding errors were reported to the main author prior to manuscript submission. The short description of the errors was emailed to the main authors and the code updates were pushed through GitHub. When needed, the code reviewers also made efforts to improve the readability of the code by breaking up long lines of code and adding comments.

Reproducibility. Reproducible code is available in the GitHub repository of the project. See Table S9 for the table of content (summary of outputs) for the main R Markdown analyses."

I cannot conduct such a review because the authors do not provide nearly enough material. I found forecast data for, presumably, Tournament 1 (https://osf.io/6wgbj/), but I found no forecast data for Tournament 2. A proper review must access all the data behind Fig S1, S2. I miss significant details for computing the MASE (what is the "training MAE" data). I cannot understand or reproduce Table 1 as the lay forecasts and realizations are not given. I cannot determine whether the statistical tests are appropriate. Etc etc.

We apologize for possible confusion. All code and data have been available on GitHub, as specified in the original manuscript. OSF was only used to provide the analytical plan previously submitted to the journal, along with the preliminary (not cleaned) version of the submitted forecasts—some forecasts had to be cross-validated with the teams for possible errors due to values being out of possible bounds, which we corrected in the version uploaded on GitHub. To avoid such confusion, we have now synced OSF and GitHub repositories, such that the same files are available on both platforms. To reiterate, each step has been cross-validated by numerous members of the Forecasting Collaborative community and we welcome further review – the GitHub repo is openly accessible.

A variance decomposition of the Tournament 1 forecast data raised concerns regarding the suitability of this data for the purpose of the study as quoted above. There are 88 teams in total, each assessing a maximum of 12 issues over a period of 12 months. Three of the teams are "revised" (1859revised, BlackSwanrevised and R4VST9 revised). That gives 85 unrevised teams, not 86 (line 204). Counting revisions there are 363 forecasts. Excluding revisions there are 363 – 7´12 = 279 not 359 (line 204). Am I looking at the right data? What am I missing?

As indicated above, the file the reviewer is referring to read was preliminary and was not cross-checked by the team. The finalized files have been listed in the GitHub repository as "wave1.scores.csv" (Tournament 1) and "wave2.scores.csv" (Tournament 2). As these files show, there were 86 teams in tournament 1. To avoid confusion, we have updated the name of the OSF file you examined earlier to "Participant_Responses for Wave 1 After Initial Data Cleaning in 2020.csv."

The number of teams assessing each of the 12 issues are shown below, ranging from 21 to 58. It is evident that the various issues are assessed by different teams, employing different methods. Does team performance take the number of assessments into account?

Yes. All analyses considered the interdependence and different number of forecasted domains via linear mixed effects models with MASE scores for domains nested in participants. As we outlined in response to Reviewer 1 above, we now provide further details about the analytical approach when introducing the results.

Are teams with few assessments choosing the better part of valor?

As we reported in the results section (see Figure 5), the number of domains teams chose to predict did not significantly contribute to (in)accuracy. If anything, teams making forecasts for more domains tended to be more accurate. This is partially due to the method forecasting teams employed: teams that relied on data tended to select more domains than specialists that relied on theory alone. This is compatible with the broader forecasting literature e.g., by Tetlock and colleagues, which shows that generalists tend to do better than specialists.

My concern with this data relative to the research question is that the forecasts are rather insensitive to time, as are many of the issues. For each issue, there is one forecast per team per month for 12 months. Below is the table for the 23 team assessments of Explicit African American Bias (eafric). Column and row averages and variances are shown. The variance in the 23´12 table equals the variance of the average + the average of the variance, computed either

column- or row-wise = 0.003295764. The variances row-wise are very small, 20 of the 23 teams are below 0.001, nearly half are below 0.0001. This means the individual teams show little month-to-month variation. On the other hand, the column variances are all above 0.001; after the first two months all are nearly equal to the total variance. The Teams explain 99.54% of the variance in this data, the Months explain 13.04% (if teams and months were independent these numbers would add to 100%). Put simply, choosing a month does not reduce the variation in forecasts, but choosing a Team does.

We appreciate this reviewer's attention to the detail when inspecting the preliminary OSF data. The reviewer 3 is right to point out that their analysis would be informative if forecasts of teams for different domains and forecasts for different months were independent. However, neither is the case in our study: Same teams could make forecasts in up to 12 domains, and month-to-months forecasts are considered to be part of a series. Because time series forecasts for consecutive months are supposed to be interdependent (forecasts for month *t* should be related to forecasts for month *t +1*), classic measures of variance can also be biased. Though beyond the scope of the present article, an unbiased metric would involve estimation of conditional variance that accounts for unique GARCH/ARIMA model of individual forecasts (such variances would not allow direct comparisons of time series with different ARIMA, though). With this caveat aside, below are summary statistics of time series variability in Tournament 1. To compare domains that rely on different units of measurement, we can look at standard deviations, and compare them to variability in ground truth time series and the previous 12 months of historical time series, to examine if scientists' forecasts are less variance than one would expect for a given time series.

domain	Median *SD*	min *SD*	max *SD*	ground truth *SD*	previous 12m *SD*
polarization	2.3	0	7.67	3.394235309	2.016504
ideolrep	1.08	0	5.55	2.286428985	1.950518
ideoldem	0.817	0	3.8	1.976928667	2.3063
posaffect	0.12	0	0.365	0.149742341	0.193748
negaffect	0.104	0	0.26	0.371527088	0.128883
lifesatisfaction	0.0521	0	1	0.035904656	0.038889
easian	0.0233	0	0.0781	0.032424575	0.064445
egend	0.0174	0	0.145	0.032715822	0.057005
eafric	0.0104	0	0.0429	0.016214714	0.02861
iasian	0.00926	0	0.127	0.010566561	0.021766
iafric	0.00413	0	0.038	0.007825543	0.003186
igend	0.00209	0	0.0614	0.005959462	0.014124

It is evident that variability across time points of the same domain varies across the teams. The reviewer is right that some forecasting teams chose to provide the same value for each of the 12

months – their *SD* for a given domain is zero. However, for many domains there are also teams that show substantial variability (more than quarter of a *SD*) in their forecasts. Critically, median forecasting variability of a given domain closely mirrors variability in the ground truth and historical time series. In other words, teams show as much (if not more) month-to-month variability as one would expect from the historical trends.

One can also clearly see this variability in the Figure S1, which depicts monthly forecasts of each team along with ground truth markers and the lowess estimate across groups. Similarly, one can see this variability in the new Figure S15, in which we plot the same data for Tournament 1 on top of the 3 years of the historical trends preceding the tournament.

Several further points are noteworthy here. First, for our metric of forecasting accuracy (MASE) it is irrelevant whether forecasts employ a months-variable strategy or a variance-free strategy concerning an estimate of a baseline + a horizontal line across 12 months (a forecasting strategy that is reasonable for domains with low historical variability). This is because MASE compares accuracy *across* 12 months rather than on a month-by-month fashion. Moreover, competitions held in the forecasting space (e.g., M competitions) are typically motivated by the need of exploring how different forecasting models perform on a given dataset (selection of which has a multitude of challenges and trade-offs, see our addition to the discussion section, inspired by feedback from Reviewer 2). In other words, the main focus of most forecasting competitions concerns the comparison of the submitted methods and how they perform in predicting societal trends under different circumstances, not on the data one has.

As the authors note: "for half of the domains the average forecasts were highly similar or nearly indistinguishable from the last historical data points provided to forecasting teams (Fig. S2)" Don't you mean fig S1?? (line 1614.).

Yes, it should be Figure S1, and we have corrected the typo.

No information is provided how the values in Figure S1 were computed (stereo-consistent?).

As we outlined in response to Reviewer 1, we have provided further details for the revised Figure S1 legend. Further, we have updated the labels to "stereotype-consistent" to avoid further confusion. We retain our description of how we calculated each ground truth and historical marker in the Methods section (now moved to the main manuscript, per journal guidelines).

There may be an important message in this data but the reader can't extract without much more information.

We hope our revision and information on GitHub provides greater clarity and avoids possible confusion about how and where to find relevant info about our methods, analytical procedures, and results, and what constituted the chief objectives of our study.

We thank the reviewer for their detailed comments.

Decision Letter, first revision:

24th November 2022

Dear Dr. Grossmann,

Thank you for submitting your revised manuscript "Insights into accuracy of social scientists' forecasts of societal change" (NATHUMBEHAV-22061597A). It has now been seen by the original referees, whose comments are included below, as well as a newly recruited reviewer with expertise in mathematical probability and forecasting, who only provided confidential comments to the editors. Although Reviewers 1 and 2 were very positive about this revision (making only minor comments for revisions prior to acceptance), Reviewer 3 was unconvinced that your data could answer the research question posed. We shared Reviewer 3's feedback with Reviewer 2 and our newly recruited Reviewer 4, both of whom felt that your data do answer the research question and your analyses are reasonable. In light of this feedback, we will be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy Reviewer 2's final comments and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Samantha Antusch

Samantha Antusch, PhD
Senior Editor
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

The authors have addressed all of the points that I raised in my prior review. Thanks for the thoughtful revisions.

Rick Klein

Reviewer #2 (Remarks to the Author):

I thank the authors for their thoughtful reply. I especially liked Figs S12, S13, S14 and S15. They have addressed my concerns, and I look forward to seeing this in print

I have a few minor questions/suggestions in no order, which I would leave to the authors discretion:

- It was a bit unclear to me what kinds of things went into the SI and which went into the methods section in the paper. I sometimes didn't know where to look.

- I believe that the yellow highlights are supposed to show changes in the manuscript, but I don't think these are complete. For example, in the accuracy across tournaments there seems to be a lot of new material (which I like), but that is not in yellow.

- Is the fact that this research happened during a very unusual period a "feature" or a "bug"? On the one hand, you could say it is a bug: how could social scientists be expected to predict in these uncertain times? On the other hand, you could say it is a feature: this should be a setting where social scientists can use their expertise better than naïve statistical models? I'm not sure myself. I wonder if the authors thing that this particular time period was one that favored social scientists or favored a naïve statistical models.

- In figure 1, I wonder if it would be clearer to label the blue square "Best naïve statistical" to clarify that the blue square is not the same approach throughout the figure, at least if I understand things correctly.

- Regarding the sentence beginning on line 713 "Given the broad societal impact . . . .the ability to predict trends in these variables would appear to be of crucial importance." I thank the authors for clarifying in their response, but I'm still a bit confused. Do they think this is of crucial importance to social scientist or policy makers (or both)? I realize that this is the last paragraph of the paper so I think it is reasonable for them to go a bit beyond what is in the paper, but I'm not really understanding this claim.

- I think paper has the potential to be read by a wide audience, and I'd recommend that the authors read through everything one last time and try to make it as clear as possible for people who are not experts in forecasting.

I'd like to thank the authors again for their work. This was a massive project, and many people will learn from it.

Reviewer #3 (Remarks to the Author):

My problem with this ms is that the data cannot possibly answer their research question about predicting social trends (line 235) because the experts and perhaps the data are practically trend free. The ms needs a fundamental re-purposing. It looks as if the issues of data availability have been addressed, but I have not undertaken a second in depth review. This ms can be published in any number of good journals. Intensive reviews for publication in the NATURE suite take a lot of time , and I don't see the value in that given the required re-purposing.

Reviewer #4
-no remarks to the Author-

Final Decision Letter:

Dear Professor Grossmann,

We are pleased to inform you that your Article "Insights into accuracy of social scientists' forecasts of societal change", has now been accepted for publication in Nature Human Behaviour.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors whose manuscript was submitted on or after January 1st, 2021, may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. IMPORTANT NOTE: Articles submitted before January 1st, 2021, are not eligible for Open Access publication. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at

rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,

Samantha Antusch

Samantha Antusch, PhD
Senior Editor
Nature Human Behaviour