FORESIGHT

The International Journal of Applied Forecasting

ARTICLES *by* FORECASTERS *for* FORECASTERS: Q1:2024

How Well Can Social Scientists Forecast Societal Change?

Join the *Foresight* readership by becoming a member of the International Institute of Forecasters
forecasters.org/foresight/

How Well Can Social Scientists Forecast Societal Change?

IGOR GROSSMANN, CHRISTOPH BERGMEIR, AND PETER SLATTERY

PREVIEW *In comprehensive research involving large-scale studies and a major forecasting competition in the social sciences, Igor Grossmann and his teams and co-authors have investigated the question of how well social scientists are able to predict societal change. They found that, in general, social scientists are not able to forecast better than laypeople or simple statistical benchmarks, and suggest better awareness and training in forecasting for social scientists.*

INTRODUCTION

It is well known that, in many areas, experts are no better at forecasting than laypersons. For example, it has been shown that stock portfolio managers and other experts are in general unable to outperform the market (Carhart, 1997; Makridakis and colleagues, 2023). However, recent studies in the social sciences suggest a more nuanced picture: researchers have found that social scientists, particularly in fields like economics and psychology, tend to make relatively accurate predictions about the replicability of laboratory experiments, especially during the COVID-19 pandemic. These findings were supported by replication projects and expert surveys, indicating that social scientists might have a predictive edge in their domains of expertise (for review, see Grossmann and colleagues, 2023). In our research, we wanted to check how successful social scientists are at predicting societal trends in some specific domains, such as gender bias, prejudice against minorities, life satisfaction and well-being, or political polarization – topics that social scientists we targeted cover routinely in graduate training and for which causal models exist.

To address this question, a group of scholars formed a Social Science Forecasting Collaborative (***predictions.uwaterloo.ca***), undertaking a range of projects on predictions social scientists provided on the news and via structured surveys during the first year of the COVID-19 pandemic (Hutcherson and colleagues, 2023). Moreover, the Forecasting Collaborative members undertook a massive research project that culminated in a research report with over 100 co-authors (The Forecasting Collaborative, 2023). In the following pages we briefly describe these research endeavors and their main findings.

In this context, it's pertinent to elaborate on the concept of expertise within social sciences, as expertise can mean many things (Collins and Evans, 2019). First, social scientists, by virtue of their extensive understanding of societal interactions and individual behaviors, possess a distinct advantage over laypeople in terms of general knowledge. Their domain-general expertise includes knowledge in areas like sociology, economics, and psychology, where they use empirical methods to dissect causal relationships in social dynamics. Second, social scientists possess expertise in specific domains like gender bias, prejudice, or well-being, bolstered by unique causal models and insights about likely variability of a particular domain of interest. In the Forecasting Collaborative, scholars therefore took a multipronged approach to expertise in quantitative social science fields, ranging from domain-general expertise compared

This article originally appeared in *Foresight*, Issue 72 (*forecasters.org/foresight*) and is made available with permission from *Foresight* and the International Institute of Forecasters.

to members of the public (without training in quantitative methods and relevant theories) to domain-specific expertise indexed by self-reported area of study, self-assessed confidence in respective domain, as well as evaluation of publication track record (cross-validated by independent raters).

Hutcherson and colleagues first examined a large text corpus of COVID-19-related news to find articles where social scientists made predictions about changes the pandemic would bring. The main finding was that social scientists would usually not resort to the use of scientific theory or data when asked to perform predictions, but would employ essentially the same heuristics and reasoning laypeople would perform. The study also included a range of large-scale surveys among social scientists and the general public about prospective judgments of the direction and magnitude of societal change over a time period at the start of the pandemic, and retrospective judgments half a year later (how much things did change). As it turned out, social scientists were largely at chance predicting the direction of change and were not better than the general public. For many domains their off-the-cuff predictions were indistinguishable from those made by the naive crowd. And curiously, accuracy did not improve when social scientists and laypeople were asked to evaluate estimates retrospectively. One important difference was that laypeople usually expressed more confidence in their forecasts than did the social scientists. Social scientists more readily admitted not being able to predict societal change well, suggesting that they were "meta-accurate." Critically, in another study in this program of research the general public perceived social scientists to be more accurate forecasters and preferred them to make predictions and provide policy recommendations during the pandemic.

In parallel to these off-the-cuff predictions, the Forecasting Collaborative hosted a formal forecasting competition. Its goal was to predict several indicators in well-studied domains with good data availability, namely subjective well-being, racial bias, ideological preferences, political polarization, and gender–career bias. Because such data was largely restricted to the U.S., the predictions zeroed in on societal change in this country. The competition was organized in two phases. The first phase asked forecasters to predict the next 12 months. The second phase started six months after the first. It therefore had six months more of data available, and it allowed the participants to resubmit their forecasts for the next six months, i.e., for the second half of the original prediction window from the first phase. In the following, we call the first

Key Points

- In many areas—the stock market, for one example—experts are no better at forecasting than laypeople or simple statistical benchmarks. We wanted to address the question of whether this is also true in the social sciences, while also identifying strategies to improve accuracy.
- In the first part, comprehensive analyses of COVID-19-related news articles and large-scale studies were performed. In the second part, we held a forecasting competition that relied on over three years of monthly data from 12 domains of societal change in human behavior to evaluate predictions during the COVID-19 pandemic.
- We found that, in most domains, domain-general expertise in social sciences (be it training in social science methods and causal theories in the general realm of forecasted domains, years of training, or academic rank) was not aligned with better forecasts compared to the general public (i.e., people without social science training) or simple statistical benchmarks, both in a competition setup and questionnaire-based studies.
- Social scientists who relied on past data, were multidisciplinary, and had some objective domain expertise and prior forecasting experience were more accurate.

phase of the competition Tournament 1 (T1), and the second phase Tournament 2 (T2).

The data available were monthly time series with 39 data points each from January 2017 to March 2020. For the second phase of the competition, this was updated to 45 data points, with data up to September 2020. As the competition was run during the COVID-19 pandemic, the prediction tasks were challenging as societal phenomena were volatile during that time.

This was the first such undertaking we are aware of, as past research in this area has concentrated on predicting geopolitical and economic events, and not societal phenomena across a range of domains in a standardized framework. It is arguably easier to predict outcomes of specific events, as they are binary compared to continuous societal trends.

The competition was open to participants from different backgrounds, with targeted announcement and concentrated recruitment efforts in social sciences, behavioral sciences, and data science. It used three statistical benchmarks and a "wisdom of the crowd" benchmark that consisted of an average of the forecasts from 802 lay persons. Teams and individuals – mostly from psychology, computer science, and economics (with over 70% possessing a doctoral degree) – received three years of historical data, were free to organize as they liked, had the opportunity to refine their forecasts in an offline setting, and were driven by the prospect of reputational gains, with their performance being publicly ranked upon the conclusion of the study.

The statistical benchmarks were a historical mean (across prior three years of monthly datapoints), a linear trend (over the same period), as well as one-step-ahead random walk (also known as a naïve forecast).

The first big question this research tried to answer was how well social scientists can forecast societal change. The answer from the forecasting competition is that, for most domains, the social scientists were not able to outperform the general public at predicting change. Furthermore, for most domains, the participants also were no better than simple statistical benchmarks. Accuracy systematically varied across domains. Positive sentiment and gender-career stereotypes were easier to forecast than the other domains included in the research. Negative sentiment and bias towards African Americans were the most difficult to forecast. The self-organized teams that performed best used past data, were multidisciplinary, and had some objectively measurable domain expertise and prior forecasting experience. In the following, we describe the setup of data and forecasting domains, and report then the main findings of the research in more detail.

THE DATA AND FORECASTING DOMAINS

By providing the datasets, the competition followed a "common task framework" enabling better comparisons among the teams. The 12 indicators that we asked participants to forecast were

- Life satisfaction
- Explicit gender–career bias
- Implicit gender–career bias
- Political polarization
- Positive affect on social media as estimated by established natural language algorithms for tracking national-level affective well-being (Schwartz and colleagues, 2016)
- Explicit Asian American bias
- Ideology Republicans
- Ideology Democrats
- Implicit Asian American bias
- Explicit African American bias
- Implicit African American bias
- Negative affect on social media.

Participants could choose any domain to forecast.

These indicators have been selected due to data availability over several years. Also, taking into account the COVID-19 pandemic, the competition used domains of societal change where social science

theories made predictions relevant to the pandemic and social isolation. In the following we report the key findings from the study.

HOW ACCURATE ARE THE COMPETITION PARTICIPANTS AT FORECASTING?

Comparing the forecasts submitted by the participants (social, behavioral, and data scientists) with an in-sample, one-step-ahead random walk using the MASE, we found that even winning teams were not better than the in-sample random walk in eight out of 12 domains. This is not very surprising as our horizons were longer than one-step ahead, and the training set was mostly prepandemic data that we assume is more predictable. Except for one team, the top performers from T1 were not under the top performers of T2. Competition participants were able to predict with significantly higher accuracy than laypeople when forecasting life satisfaction, polarization, and explicit and implicit gender–career bias, but no better than the laypeople benchmark in the other eight domains.

COMPARISON AGAINST NAIVE STATISTICAL BENCHMARKS

First, we compared the participants' forecasts against all three statistical benchmarks. We calculated benchmark/forecast ratio scores, where values greater than one indicate the competition participants were more accurate than the benchmark. In T1, the participants were significantly better than all of the three benchmarks in only one out of the 12 domains. In T2, the scientific forecasts were somewhat better than mean and random-walk forecasts, outperforming the benchmarks now in five of the 12 domains.

Next, we compared against the forecast combination (average) of the three benchmarks. Here, in T1, overall, the competition participants struggled to achieve higher accuracy than the average of the three benchmarks. In most domains, one or more of the benchmarks was at least as good as social scientists' forecasts.

WHICH DOMAINS ARE HARDER TO PREDICT?

We found that some societal trends were much harder to predict than others. Predictions for political support of the Republican party had the lowest accuracy, while forecasts for explicit gender–career bias and positive sentiment on social media had the highest accuracy. Complexity in historical data, namely more variability in the past data, also made the data more difficult to forecast. Lastly, we checked for influences on whether the data comes from very reliable sources or not. We found that this did not affect forecasting accuracy.

COMPARISONS OF ACCURACY ACROSS TOURNAMENTS

We found that accuracy was lower in T1 compared with T2. This can be for various reasons. Our first idea was that the teams may have changed. However, when controlling for team characteristics, the differences remained.

One prominent difference is that for T1, participants had to predict the next 12 months, whereas for T2, only six months ahead were required. However, we found that in T1, error in the first six months was greater than in the second six months. Participants underpredicted in general in T1, and underpredicted more in the first six months than the second six months. Many domains showed unusual shifts because of the pandemic in the first months, but later returned to the historical baseline. Thus, an explanation could be that T2 had higher accuracy as data from the pandemic was available in T2.

CONSISTENCY IN FORECASTING

The forecasts were consistent in the evaluation we performed where odd months achieved very similar accuracies to even months (a sanity check to establish individual systematicity and reliability in forecasting accuracy – it was high even though accuracy itself was low; in other words, the forecasts people provided were not ad hoc), and accuracy in T1 was associated with higher accuracy in T2.

Harder-to-predict domains from T1 remained hard to predict in T2, with one notable exception being bias against African Americans, which became easier to predict. We speculate that this may be a result of police brutality and racial inequality protests and resurgence of the Black Lives Matter movement and the associated higher racial awareness in the aftermath of George Floyd's death.

WHICH STRATEGIES AND TEAM CHARACTERISTICS PROMOTED ACCURACY?

Teams used strategies based on historical data, based only on theories and intuition, or on both in a hybrid fashion. Forecasts that used data were more accurate than forecasts that did not. Furthermore, data-free predictions were not more accurate than laypeople's predictions. Simpler forecasting models resulted in significantly higher accuracy than more complex models. The subjective confidence of the participants in their forecasts was not related to their achieved accuracy. However, teams with more expertise made more accurate forecasts, suggesting that there is some value in domain-specific expertise, which was measured as published papers by team members in the forecasting domain.

Furthermore, through the way the competition was organized, participants could effectively update their forecasts from T1 in T2. This made the forecasts significantly more accurate than without the updating. However, they were not more accurate than the forecasts from teams that had entered the competition directly in T2.

WHICH STRATEGIES WERE THE MOST EFFECTIVE TO PRODUCE GOOD FORECASTS?

Participants who relied on prior data to generate their forecasts achieved higher accuracy than laypeople, and were also more likely to be among the top-performing teams. Conversely, persons who indicated that they chiefly relied on specific theories/causal models in predicting a particular domain were not more accurate than others. Forecasting experience and subject matter expertise also contributed to better performance in the competition (albeit the gains were modest for the latter). In particular, publication track record contributed more to achieve better accuracy than subjective (over) confidence in the forecast accuracy or domain expertise.

WHY DID THE SOCIAL SCIENCE FORECASTERS PERFORM SO BADLY?

A possible explanation could be that our incentives were not enough, as no cash prizes were involved in the competition. However, evidence of monetary incentives performing better than reputational incentives like those used in the tournament (by announcing the winners) remains the subject of debate. Furthermore, social scientists often deal with phenomena that have small effect sizes, and are overestimated in the literature (Open Science Collaboration, 2015), as effects can be smaller in the real world compared with the lab studies that social scientists perform. Moreover, social scientists in such fields as psychology and organizational behavior tend to research individuals and small groups, which may not readily scale to whole societies. Also, training in predictive modeling is not part of many social sciences curricula, as theories are usually developed with explanatory power and not predictive power in mind. Lastly, theories and models developed pre-COVID-19 may not be readily applicable during and after the pandemic, though this was taken into account during data selection, as outlined before.

HOW CAN SOCIAL SCIENTISTS BECOME BETTER FORECASTERS?

Not all social science theories may be testable through forecasting, but this could be an important standard criterion to evaluate a theory, where applicable. Using insights from the time series literature can lead to better societal models and insights, and thus, social scientists would benefit from better forecasting skills. For example, by testing whether societal trends are deterministic or stochastic,

researchers can also determine whether attempting to develop a causal theory is worthwhile or whether the model can be explained in terms of stochastic processes. To better understand the characteristics of the societal trend, social scientists may also benefit from performing seasonal decompositions, or tests for nonstationarity of their data. Furthermore, practicing forecasting and monitoring accuracy more regularly, through services such as Metaculus (metaculus.com) or Prediction Book (predictionbook.com), may also help to improve capabilities.

Some of the variables included in the competition have broad impacts on societies and predicting them accurately is important for policymaking. As such, forecasting in these domains is important. However, social scientists currently seem to have quite limited capabilities in this area when taken as a whole, with only one marker (publication track record) yielding some modest albeit significant advantage in the forecasting tournament. Most other markers of expertise were unrelated to their forecasting performance in the tournament, and none of the expertise markers were related to forecasting accuracy in prior large-scale surveys. As such, we should strive to improve their forecasting skills.

SUMMARY

We have summarized our recommendations in **Table 1**.

Table 1. Summary of Recommendations

AUDIENCE	RECOMMENDATION
Forecasting practitioners	Expect complexity and variability in historical data to increase forecasting difficulty.
	Use relevant past time series data when forecasting.
	Try to work in diverse teams with different skill sets when forecasting.
	Be aware of the large uncertainties the future holds and be cautious of overconfidence.
Forecasting researchers	Consider using different types of incentives (e.g., monetary, reputational) in competitions to ensure "skin in the game" by participants.
Social, behavioral and data scientists	Since post-hoc explanation and forecasting can be different goals that require different approaches, consider using forecasting techniques when forecasting.
	For example, evaluate whether the trend is stochastic vs. deterministic and consider how to separate the time series process due to seasonality and nonstationarity from the deterministic impact of exogenous factors.
General public	Trust experts most in the narrow field they are expert in, by relying on more objective markers of expertise such as publications in a given domain (rather than subjective confidence ratings).

Igor Grossmann is a Professor of Psychology at the University of Waterloo. Using computational, experimental, and psychometric methods, he researches societal change, expert judgment accuracy, and wisdom. Igor has a PhD in psychology from the University of Michigan, received awards from APA, APS, SPSP, and is the College Member of the Royal Society of Canada. He founded the Forecasting Collaborative, WorldafterCovid.info project, and co-hosts the "On Wisdom" podcast.

igrossma@uwaterloo.ca

Christoph Bergmeir is a María Zambrano Senior Fellow in the Department of Computer Science and Artificial Intelligence at University of Granada, Spain. Before this he was a Visiting Research Data Scientist at Meta Inc. in the U.S., and a Senior Lecturer at Monash University, Australia. Christoph holds a PhD in computer science from the University of Granada and an MSc degree in computer science from the University of Ulm (Germany). He has worked in forecasting for capacity planning, sustainable energy, and supply chain.

bergmeir@ugr.es

Peter Slattery is a Visiting Researcher at MIT FutureTech, where he explores trends in AI adoption, and AI risk awareness and preparedness. Peter has a PhD in information systems from the University of New South Wales and an MBS in information systems from University College Cork. He founded Ready Research and the EA Behavioural Science Newsletter.

pslat@mit.edu

REFERENCES

Carhart, M.M. (1997). On Persistence in Mutual Fund Performance, *The Journal of Finance* 52(1), 57–82.

Collins, H. & Evans, R. (2019). *Rethinking Expertise*, University of Chicago Press.

The Forecasting Collaborative (2023). Insights into the Accuracy of Social Scientists' Forecasts of Societal Change, *Nature Human Behaviour*, 7(4), 484–501.

Grossmann, I., Varnum, M.E., Hutcherson, C.A. & Mandel, D.R. (2023). When Expert Predictions Fail, *Trends in Cognitive Sciences*.

Hutcherson, C.A. et al. (2023). On the Accuracy, Media Representation, and Public Perception of Psychological Scientists' Judgments of Societal Change, *American Psychologist*, 78(8), 968-981.

Makridakis, S. et al. (2023). The M6 Forecasting Competition: Bridging the Gap between Forecasting and Investment Decisions, arXiv preprint arXiv:2310.13357.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science, *Science*, 349(6251), aac4716.

Schwartz, H.A. et al. (2016). Predicting Individual Well-being through the Language of Social Media, *Biocomputing 2016*: Proceedings of the Pacific Symposium, 516-527.