Reciprocal Scoring: A Method for Forecasting Unanswerable Questions[‡]

Ezra Karger[*, a] Joshua Teperowski Monrad[*, b] Barbara Mellers[c] Philip E. Tetlock[†, c]

[a] *Federal Reserve Bank of Chicago*
[b] *Future of Humanity Institute, Oxford University*
[c] *Wharton School, University of Pennsylvania*

October 31, 2021

We propose an elicitation method, Reciprocal Scoring (RS), that challenges forecasters to predict the forecasts of other forecasters. Two studies show how RS can generate accurate forecasts of otherwise unanswerable questions. Study 1 establishes the epistemic credibility of RS: forecasters randomly assigned to use RS were as accurate as forecasters predicting objectively resolvable outcomes using a proper scoring rule—and both groups were more accurate than a control group that felt accountable to neither intersubjective RS metrics nor objective metrics. Study 2 establishes the practical value of RS. We ask highly accurate forecasters to predict each other's forecasts of the effect of government policies on COVID-19 mortality, yielding a real-time ranking of the expected effectiveness of pandemic-containment policies. As in Study 1, RS forecasters converged but in this case on policy recommendations that stand up to scrutiny, even with the benefit of hindsight. The core contribution of RS is its power to create accountability for accuracy in policy debates that have long been stalemated by the absence of accountability.

Keywords: forecasting tournaments, causal inference, policy evaluation, elicitation, COVID-19

[‡]The authors thank Zachary Jacobs, Yunzi (Louise) Lu, and Raman Thadani for extensive research assistance; Amory Bennett and Charlie Bailey for customizing the forecasting platform; Kaitlyn Coffee for operational support; and Daniel Kahneman, Pavel Atanasov, and Chris Karvetski for insightful discussions of our work. This work was supported by Open Philanthropy and the Intelligence Advanced Research Projects Activity (IARPA), contract number 140D0419C0049. The U.S. Government is authorized to distribute reprints notwithstanding any copyright annotation. The views expressed here are those of the authors, not of IARPA or the U.S. Government. And the views expressed also do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

[*]Equal contribution.
[†]Corresponding author.

Introduction

Subjective-probability forecasting tournaments have played a largely peripheral role in public policy debates. One reason is that such tournaments focus on trend-tracking, with few efforts to address how trends will vary conditional on policy interventions under debate. The emphasis, especially in the first generation of geopolitical tournaments, has been less on improving public policy than on improving human judgment by narrowing the gap between subjective probabilities and objective reality, as gauged by proper scoring rule functions (Mellers et al., 2015; Tetlock, 2005; Tetlock & Gardner, 2016; Winkler & Jose, 2011).

Of course, researchers have made policy-relevant discoveries. For instance, we have learned that crowds of talented engaged amateurs could outperform public health professionals on an array of near-term trend COVID-19 questions (Metaculus, 2020), extending past work on the limitations of expert judgment (Tetlock, 2005). But the efforts suffer from the same rigor-relevance gap that plagued first-generation subjective-probability tournaments from their inception. The mass media often find tournaments engaging: How many will die from COVID-19? Can talented amateurs beat the experts again? How sustainable is "superforecaster" performance? But the answers tend to be of dubious relevance to policy-makers wrestling with resource constraints and complex trade-offs. Indeed, there is a real risk of first-generation tournaments degrading signal-to-noise ratios in policy debates by publishing misleading horse-race statistics. The relative neglect of policy evaluation is partly driven by the fact that questions about the causal impact of policies can rarely be objectively 'resolved,' making ex-post evaluations of accuracy impossible. The elusiveness of objective answers to the questions we most urgently need answered deprives forecasting tournaments – such as those hosted by Good Judgment Open, Metaculus, and prediction markets – of the gold-standard accuracy metrics that researchers want for identifying skilled forecasters and incentivizing analytic effort. A similar problem plagues forecasting of long-term futures, given the difficulty of evaluating and rewarding accurate forecasts when objectively correct answers may not be available for decades.

This article proposes a method for achieving greater policy relevance at negligible cost in rigor.

The method – Reciprocal Scoring – incentivizes independent panels of highly accurate forecasters to predict each other's predictions, providing high-quality forecasts of unresolvable questions, such as estimates of counterfactual policy effects and long-run forecasts. We present a randomized experiment that compares forecasts elicited from Reciprocal Scoring to forecasts elicited using standard Brier scoring. We show that the new methods elicit forecasts of comparable accuracy. We also put the method to work by asking panels of "superforecasters" to predict each other's predictions of the causal impact of 13 distinct U.S government policies on COVID-19 mortality in 2020 and 2021. We use the results to rank policy responses to COVID-19 by their causal effect on mortality. These studies are proof of concept that human-judgment forecasting can generate timely, plausible, and reliable rankings of the effectiveness of competing policy options—and that it is possible to close the rigor-relevance gap.

Conceptual challenges for predicting causal effects

To achieve policy relevance, subjective-probability forecasting tournaments will have to emphasize the quality of the questions as much as the accuracy of the answers (Tetlock, 2017). Unfortunately, there is less scientific consensus on quality-of-question metrics than there is on accuracy-of-answer metrics, where we can draw on a sophisticated literature on proper scoring rules that incentivize truthful reporting (Jose & Winkler, 2011). What counts as a good question depends on the goals of inquiry. If we want to identify talent using item-response theory, we need a battery of questions that efficiently distinguishes better from worse forecasters (Bo et al., 2017; Merkle et al., 2017). And if we want to evaluate the cost-effectiveness of life-saving interventions, we need questions that efficiently differentiate policy options with higher or lower expected values (Nelson, 2005).

Consider the problem of advising the U.S. Government on time-sensitive issues concerning the COVID-19 pandemic. An intuitively appealing idea would be to break each issue into pairs of conditional-forecasting questions, like: (1) "Estimate the U.S. death toll from COVID-19 in 2021,

if the U.S. Government mandates use of facemasks for at least three months during 2020?" and (2) "Estimate the U.S. death toll from COVID-19 in 2020, if the U.S. Government *does not* mandate facemask usage for at least three months during 2021?"

Although it may be tempting to interpret differences between answers to (1) and (2) as estimates of the causal impact of a mask mandate on mortality, it is a temptation we should resist. Causality between COVID-19 severity and mandates could well be bidirectional. Mandates could drive down mortality, and mortality could drive up support for mask mandates. A forecaster might conclude that the mask mandate is likeliest in high-mortality scenarios and predict *higher* mortality rates conditional on the policy–even if the forecaster is convinced that the mandate would reduce deaths.

Policy-evaluation tournaments using Reciprocal Scoring preempt such confusion and cut to the core causal and counterfactual questions that policy-makers need answered. Such tournaments can ask forecasters paired conditional questions about the outcome of interest absent the policy intervention, and about the same outcome, conditional on the policy being implemented immediately, holding all else equal in the immediate short-run.

Targeting the impact of immediately implemented policies cuts through the causal ambiguities of conventionally paired conditional questions. Forecasters no longer need to ponder the political circumstances under which a mask mandate is likely to pass and can focus exclusively on the epidemiological challenges of estimating the direct effect of the mask mandate on mortality, assuming immediate implementation and holding all else equal. And by eliciting a baseline forecast about the outcome of interest absent the intervention, we can treat the difference between these two forecasts as a direct estimate of the effect of the mask policy.

This approach to estimating causation can be combined with well-specified interventions to evaluate impacts on complex events. Imagine estimating global vulnerability to another pandemic. Let's define event Y as "deaths from a global pandemic exceed 1 million in 2023" and action X as "the U.S. doubles annual funding for the WHO during 2021 and 2022." In a tournament, we can estimate how policy X affects the probability of Y by asking well-calibrated forecasters to estimate:

1) The unconditional probability of Y.

2) The probability of Y, assuming immediate implementation of policy X.

The difference between (1) and (2) yields, in theory, a direct estimate of the causal effect of implementing X immediately on p(Y).[1] Of course, "in theory" does not imply "in practice."

Using Reciprocal Scoring to elicit rigorous forecasts

The practical challenge becomes: how can we elicit high-effort truthful and accurate beliefs when we lose the central advantage of forecasting tournaments, the objective resolvability of questions? Reciprocal Scoring answers this question with a new accuracy metric: recruit forecasters with strong track records who are presumptively motivated by epistemic goals, divide them into independent groups, and then ask each forecaster to predict the median of the other group's forecasts.

There are numerous instructive precedents for this peer prediction approach in a large literature on judgment aggregation, signal elicitation, output agreement mechanisms, and choice-matching. These range from Keynes' (1936) famous beauty-contest thought experiment, the large literature on eliciting beliefs from participants in a game where the true state of the world is unknown (Myatt & Wallace, 2012), work on proxy scoring rules (Liu et al., 2020; Witkowski et al., 2017) and Prelec's research program on Bayesian truth serum (Cvitanić et al., 2019; Frank et al., 2017; McCoy & Prelec, 2017; Prelec, 2004; Prelec et al., 2017). For instance, Cvitanić et al. (2019) propose

[1]Note that the unconditional probability of Y may include some probability that X is implemented in the future. The goal is to estimate the causal effect of immediate implementation of X on Y, relative to the current world, where X may have some non-zero chance of being implemented. Another version of this tournament structure would elicit the unconditional probability of Y, the probability of Y assuming immediate implementation of policy X, and the probability of Y assuming no immediate implementation of policy X. This would fully differentiate the effect of X on Y from the effect of not implementing X on Y. It is important, both in our application and in this alternative structure that the implementation of policy X be immediate. If we instead attempt to forecast the probability of Y given implementation of policy X in one year, then that year may itself affect the relationship between X and Y, which complicates interpretation. For example, if we ask participants in March 2020, at the onset of the Covid-19 pandemic in the U.S., to forecast the effect of a mask mandate implemented in March 2021 on Covid-19 deaths in 2021, then forecasters must simultaneously forecast the effects of a mask mandate in March 2021 on Covid-19 deaths in 2021 under each possible state of the world in March 2021, aggregating those forecasts into an overall estimated causal effect. In this paper, we avoid this by asking forecasters to predict immediate policy effects in Study 2.

a choice-matching method for eliciting subjective beliefs by posing a multiple-choice question. Suppose the respondent picks option (a). The authors then ask each respondent to predict the distribution of forecasts of other respondents and give respondents a score that is a function of the predictions of those who also answered (a) on the initial question. Forecasters are scored according to the distance between their predicted distribution and the actual distribution of forecasts. With the aid of certain assumptions (e.g., that all forecasters have a common prior on the number of forecasters of each type), Cvitanić et al. show that this method elicits truthful responses to questions that are not objectively resolvable. Building on Prelec's (2004) work on Bayesian truth serum, Frank et al. (2017) have also shown that Prelec's method improves honesty in online surveys.

In addition to the literature described above, we want to highlight two prior studies that explore peer prediction using a similar framework where each forecaster predicts the forecast of peers. Waggoner and Chen (2014) introduce a theoretical model of output agreement mechanisms, where two forecasters are asked a question and given a reward monotonically decreasing in the distance between the two responses. Waggoner and Chen show that the equilibrium of this game elicits common knowledge from forecasters. Relatedly, Court et al. (2018) empirically test the accuracy of box office revenue forecasts elicited using a 'guessing the guesses' mechanism and a small (N=167) empirical study without a control group. We expand on these ideas in four important ways. First, we provide the first large-scale empirical test of this type of peer-prediction mechanism, showing that Reciprocal Scoring elicits forecasts that are both statistically indistinguishable from forecasts elicited using a proper scoring rule and statistically distinguishable from forecasts elicited from a control group without an accuracy incentive; second, we extend these ideas by forecasting unresolvable policy counterfactuals, ranking pandemic-mitigation policies in the midst of the Covid-19 pandemic and evaluating our ranking relative to retrospective empirical analysis; third, we elicit these forecasts from highly-accurate forecasters, making us confident that our elicited beliefs represent high-quality estimates of those causal policy effects; and fourth, we provide a blueprint for how policy makers can use Reciprocal Scoring to improve decision-making in other contexts where causal policy effects are important, but unknown.

To understand why Reciprocal Scoring incentivizes forecasters to submit their best guesses about the truth, we need to put ourselves in their shoes. Both the monetary and reputational incentives encourage each forecaster to figure out what other forecasters will do and why. And in our framework, which relies on high-credibility forecasters with strong track records, each forecaster expects the others to submit high-quality forecasts. What, then, is the optimal strategy for a forecaster? An equilibrium in this tournament is a set of individual forecasts such that no forecaster can earn a larger prize by changing his or her forecast, holding others' forecasts constant. One equilibrium (the preferred one, from a policy perspective) is if everyone understands the format, carefully researches the questions, and estimates what 'high-quality' forecasters would forecast by submitting their best guesses about likely impacts of policies.

To operationalize the process, suppose we ask a question that will remain unresolvable during our lifetimes: how many people will be killed by pandemics in the next century? Will forecasters react to the ground rules and incentives as intended or look for lower-effort "gaming-the-system" solutions?

Consider a stylized example with three forecasters, Anna, Bob, and Charlie, who each must submit forecasts without knowing the others' strategy. For simplicity, assume each could submit a low-effort forecast of 0, a medium-effort forecast of 100 million (perhaps the result of a quick Google search), or a high-effort forecast of 500 million after serious modeling. Also, assume that our accuracy prize is equal to a large dollar amount if the three forecasters choose the same forecast, and 0 otherwise.[2]

Three simple, pure-strategy equilibria are:

1. Anna, Bob, and Charlie all submit the low-effort forecast of 0
2. Anna, Bob, and Charlie all submit the medium-effort forecast of 100 million
3. Anna, Bob, and Charlie all submit the high-effort forecast of 500 million

Importantly, the tournament organizer and forecasters do not know that the high-effort (likeliest to be accurate) forecast is 500 million—after all, that is the answer the organizer is seeking. And

[2]This is just one easy-to-model prize function that is (weakly) increasing in accuracy relative to your peers.

in all three equilibria, no forecaster has an incentive to stray because doing so would reduce her prize from a positive value to zero.[3] So, there are grounds to worry this method will yield sub-par equilibria that will dissuade sponsors from launching such exercises in the first place.

There are however methods of preventing sub-par equilibria, with the aid of modest, empirically verifiable assumptions. Suppose that at least one forecaster, Anna, is a good-faith forecaster who always invests high effort. Suppose further that Bob and Charlie know there is at least one high-effort forecaster in this tournament and the expected prize is large enough to elicit high effort from the good-faith forecaster. Lastly, suppose that conditional on implementing a high-effort search for the answer, all forecasters believe that the other forecasters will converge on the same answer (a common prior). In this case, the only pure equilibrium is the high-effort equilibrium.[4] Any other forecast will yield a lower expected prize.[5]

Of course, Reciprocal Scoring is not knave-proof. We must address at least four potential deviations from the desired equilibrium:

(a) Forecasters might submit forecasts that do not correspond to their true beliefs because they expect their colleagues to be biased. Suppose a forecaster suspects that others are too optimistic about future pandemics—and adjusts her forecasts accordingly, reporting something other than her true beliefs. But if the tournament consists of forecasters known to have strong track records, then assuming a biased response from others ceases to be a prize-maximizing strategy. For example, in several large-scale tournaments funded by IARPA, "superforecasters" were surprisingly well-

[3] Assume that the prize pool is between \$0 and \$100 for each forecaster and is not a function of the prizes that other forecasters receive (which would complicate this tournament).

[4] But are there any mixed strategy equilibria? The answer is no because Anna is assumed to forecast the high-effort equilibrium in all states of the world. Bob and Charlie cannot improve their monetary prize by using any form of mixed strategy—this would net them a monetary reward of zero in some states of the world and they would be better off only forecasting the high-effort equilibrium in all states of the world.

[5] This basic model generalizes quite easily to $N > 3$ forecasters. Because of our simple scoring rule and the assumption that Anna always is known to submit a high-effort forecast, the only equilibrium is for all other forecasters to do so as well, even if there are 10, 100, or 1 million other forecasters. Importantly, this is true even if non-Anna forecasters are unsure about the identity of the high-effort forecaster. We propose a slightly modified version of this setup for actual implementation: divide all forecasters into two independent teams and give out a prize to each forecaster inversely proportional to the squared distance between their individual forecast and the median forecast of forecasters on the other team. Behaviorally, this will lead to more effort than the sharper scoring rule above. And relying on well-organized teams that value open dissent, instead of individual forecasters, will ensure that forecasters produce higher-quality forecasts (Tetlock & Gardner, 2015)

calibrated on a broad set of questions: from standard global events (economic output, elections, pandemics) to predicting patterns of outcomes in simulations of complex social systems (Tetlock & Gardner, 2016).

(b) Forecasters might submit forecasts that deviate from their true beliefs because they think they have access to private information unavailable to the other team. Suppose a forecaster knew about an article in an uncommon language that had not been translated and that others would be unlikely to see. Even if this information did affect the forecaster's private beliefs, she would be incentivized to ignore it if she suspected others would be in the dark. The remedy here is for tournaments to deploy large teams with free internal flows of information, large enough that the likelihood of private useful information drops fast. Suppose the forecaster puts an independent 10% chance on another discovering the information. In a 100-person tournament with two 50-person teams, this yields only a $0.90^{50} = 0.5\%$ chance that the other team would miss the information, which gives the forecaster strong reasons to incorporate the information into her reported forecast.

(c) A third concern is that Reciprocal Scoring will permit laziness. Suboptimal effort is hard to police but there are counter-measures, including (1) working with forecasters who value their reputations for integrity and rigor among teammates; (2) requiring all forecasters to make regular visible contributions to collective analytic discussions; (3) creating a within-team competition to generate rationales that external judges will evaluate on dimensions such as creativity, logical rigor, and empirical accuracy. Given the performance of top forecaster teams in first-generation tournaments, there are grounds for optimism about the feasibility of nudging teams toward progressively higher-effort equilibria.

(d) One last worry is that forecasters will game the system. This has not been a problem in first-generation tournaments because the resolution criteria are objective. But Reciprocal Scoring could open the door to cheating. If a forecaster can find a friend on the other team, the two could coordinate to maximize their prizes—skipping the hard analytic work of forecasting. Designers of reciprocal-scoring tournaments can prevent such cheating by: (1) tightening anonymity of team membership; (2) monitoring how forecasters are thinking about problems to detect bias or sloppi-

ness; (3) expanding the set of teams beyond two and evaluating forecasters based on the distance between their forecasts and aggregate forecasts of the other teams — or on the distance between their forecast and randomly selected individuals on other teams.

To formalize the theoretical basis for Reciprocal Scoring, consider the following model, drawing from the framework of Myatt & Wallace (2012). A tournament organizer wants to know the best possible forecast (an unknown real number θ) about an event. θ could represent a forecast of coronavirus deaths in the world or the global population in 2030. A continuum of forecasters perform the following steps:

1. Each forecaster pays a cost C(e) to exert effort 'e' and obtains an independent draw of information θ_i about the world from a normal distribution with mean θ and variance $\frac{1}{e}$.[6] The signal is unbiased because its distribution is centered on the true value θ. So, the more effort forecasters exert, the more likely it is their draw from this information source will yield a signal closer to the truth. The cost function C(e) is assumed to be increasing, convex, and differentiable. Intuitively, the cost function could represent cost from the time spent gathering information.

2. The forecaster submits a forecast 'f' to the tournament organizer.

3. The forecaster receives unobserved utility $U = -\alpha * (f-\theta)^2 - \pi * (f-\bar{f})^2 - C(e)$. In this utility function, $\alpha > 0$, is a fixed (unobserved) parameter that reflects the intrinsic importance of the truth to each forecaster, and $\pi > 0$ is a prize that the tournament organizer pays to participants as a function of the difference between their forecast and the average forecast of other forecasters; $\bar{f}$ is the average of all forecasters' forecasts, as received by the tournament organizer, in the style of Reciprocal Scoring.[7]

In this model, a forecaster's best option after selecting effort 'e', in response to any signal θ_i is to submit a forecast 'f' that maximizes her own utility, which is done by choosing 'f' such

[6] Following Myatt & Wallace, 2012, we assume that θ has an improper prior, with the same prior probability given to any value of θ.

[7] In our empirical analyses of Reciprocal Scoring we use the median and not the average of other forecasters' forecasts to avoid a case where a small number of outliers affects our results.

that the first derivative of the utility function with respect to f is zero–in other words, $\alpha * (f - E[\theta|\theta_i]) + \pi * (f - E[\bar{f}|\theta_i]) = 0$. The forecaster accomplishes this by submitting a forecast f equal to $\alpha * E[\theta|\theta_i] + \pi * E[\bar{f}|\theta_i]$. In other words, the forecaster will forecast a weighted average of her best estimate of the value of θ given her personal signal θ_i and her expectation of what the other forecasters will forecast (again, given her personal signal θ_i). Because each forecaster only receives one unbiased signal about the true value of the world (θ_i) and the other forecasters do as well, this weighted average corresponds to submitting a forecast of her own signal, θ_i.

In this signaling model, a forecaster cares at least marginally about two quantities: how close her forecast is to the truth (intrinsic motivation) and how close her forecast is to the average forecasts of other forecasters–a quantity we incentivize directly using Reciprocal Scoring and a prize π. This incentivizes the forecaster to submit her best estimate of the truth.

To go a step further–how much effort will a forecaster exert? She will exert effort to maximize her own utility, which involves internalizing a tradeoff between the increased cost of obtaining a precise signal (channeled through C(e)) and the increased payoff of submitting a forecast that is closer to both the unobserved true value (θ) and to the average forecast submitted by other forecasters. In the absence of Reciprocal Scoring, where forecasters care only about the distance between their forecast and θ, tournament organizers have no way of incentivizing effort because θ is unobserved. In this model, where Reciprocal Scoring creates an incentive for forecasters to report forecasts similar to other unbiased forecasters, the tournament organizer can increase the accuracy of forecasts by reducing the cost of effort or by increasing the payoff from exerting effort (which is equivalent to increasing π). Practically, this could involve offering large prizes for forecaster effort. Because each forecaster's signal is independent conditional on the true value of theta, the median or average of all forecasters' forecasts is more precise than any individual forecast, and the precision of the median or average of all forecasts is increasing in e (and therefore increasing in the value of the prize). So, a tournament organizer with large incentives to maximize accuracy will use prizes to encourage a high level of individual effort from forecasters.

Empirical Work

We report two studies of Reciprocal Scoring. The first study is a randomized experiment that tests whether we have to pay a price in accuracy for the policy-relevance advantages of Reciprocal Scoring. We show that this method yields predictions on resolvable short-term questions that are as accurate as those of conventional Brier-scoring methods of assessment—and both conditions outperform a low-epistemic-accountability control group. The second study focuses on the capacity of the new method to deliver insights into policies that first-generation tournaments are incapable of delivering, with special emphasis on COVID-19.

Study 1: Reciprocal Scoring with nonparametric probability elicitation

Methods and materials

We preregistered the design, sampling plan, exclusion criteria, measures, and analysis plan at aspredicted.org (#66990).[8] A total of 1,479 participants completed an online survey administered through Prolific on June 4-5, 2021. We asked participants to forecast short-term outcomes on June 30th relating to a range of topics, including COVID-19 cases, deaths, and vaccinations, the weather, the stock market, and oil prices (see Appendix A for an overview of the questions). For each question, we asked participants to forecast the 5th, 25th, 50th, 75th, 95th percentiles of the outcome. For example, when asked to forecast the number of new COVID-19 cases on a future date, participants specified values X to complete sentences of the form "There is a 5% chance that the number of cases will be less than X" and "There is a 25% chance that the number of cases will be less than X",... Participants recorded their answers on a set of sliding scales for each of the respective percentiles. Appendix H shows an example of the question structure.

We excluded 195 participants based on pre-registered exclusion criteria, such as failing comprehension and attention checks (N = 67), completing the survey in less than 15 minutes (N = 56),

[8] We also performed a study with an independent set of questions and forecasters almost a year earlier to pre-test these methods. That study, while underpowered, also showed that Reciprocal Scoring elicited forecasts as accurate as forecasts elicited using Brier scores (no statistically significant difference between the two groups).

or providing logically impossible, non-monotonic forecasts on at least two questions (N = 72), leaving us with a final sample of 1,284 participants. In addition, we recruited 17 highly accurate forecasters – "Superforecasters" identified during the IARPA-funded Aggregative Contingent Estimation (ACE) forecasting tournament – to complete the survey. Given each participant provided five distributional forecasts (the 5th, 25th, 50th, 75th, and 95th percentile forecasts) for each of ten questions, our analysis draws on a total of 59,600 forecasts after exclusions. (See Appendix A for more detail, including demographics.) We paid all Prolific participants $5, and Superforecasters were paid $25.

Before making their forecasts, participants were randomly assigned to one of three conditions, corresponding to three approaches to evaluating forecasts: Brier scoring (N = 423), Reciprocal Scoring (N = 428), and a Control condition where forecasts were told they would not be scored (N = 433). Finally, the 17 Superforecasters were automatically assigned a separate condition identical to the Brier condition. In the Brier condition, participants were informed that the five most accurate forecasters would receive a monetary bonus of $100. Accuracy would be measured using a Brier score, and they were provided an intuitive visual explanation of the scoring method (i.e., the closer their forecast to the actual outcome, the better the score). Similarly, in the Superforecaster condition, participants were informed that the five participants with the best Brier scores would win a monetary prize. In the Reciprocal Scoring condition, participants were also informed of the $100 prize for the five most accurate forecasters but were told their accuracy would be determined according to Reciprocal Scoring, which was also explained in detailed but intuitive terms: the closer their forecasts were to those of an expert panel of Superforecasters, the better their score. In the Control condition, participants were told that five winners of $100 would be randomly drawn from all who completed the survey (no mention of accuracy mattering). Appendix I shows the training material that each experimental condition received describing the elicitation mechanism and scoring rules. In addition to the forecasting accuracy, we measured how much time participants spent on the survey and how many information sources they consulted.

Our primary dependent variable (DV1) was the individual-level accuracy of respondents in each

condition. For each of the 1,301 forecasters on each of 10 questions, we calculated their accuracy as the squared difference between each forecasters' 50th percentile forecast for that quantity (the number X such that the forecaster believes there is a 50% chance the value will be above X and a 50% chance the value will be at or below X) and the true value of the forecasted quantity on June 30th (e.g., the actual temperature or the number of COVID-19 vaccinations in some location on that date). We standardized this measure on each question by subtracting the control group's average accuracy on that question and dividing it by the control group's standard deviation of accuracy. The primary measure of individual-level accuracy was then the average (across all 10 questions) of the standardized question-level accuracy, which we then re-standardized by subtracting the mean of the control group's accuracy across participants and dividing by the mean of the control group's standard deviation of accuracy across participants. Appendix B shows question-level median forecasts by experimental group.

As our secondary dependent variable (DV2), we used the full distributional set of 5th, 25th, 50th, 75th, and 95th percentile forecasts from each question. We calculated this individual-level measure of accuracy for each question as the squared difference between each percentile forecast and the true value on June 30th.[9] We then standardized this measure of accuracy by subtracting the control group's average accuracy on that question and dividing by the control group's standard deviation of accuracy. We then average question-level accuracy to the forecaster level and lastly, re-standardize the person-level accuracy measure by subtracting the control group's average accuracy on that question and dividing by the control group's standard deviation of accuracy.

Our pre-registered analysis compared the median of our primary accuracy measure in the Brier and Reciprocal Scoring conditions, to test whether, as predicted, they would produce similar levels of accuracy. We also compared the secondary accuracy measure in the Brier and Reciprocal Scoring conditions—as well as the Reciprocal Scoring condition to both the Brier and Control conditions. Finally, we compared all conditions to the accuracy of the Superforecaster condition.

[9] For example, if a forecaster predicted that the temperature on June 30th in New York City would be under {60,70,80,90,100} degrees with respective probabilities {5%, 25%, 50%, 75%, 95%}, and the true value on June 30th was 75 degrees, then the individual's unstandardized accuracy measure on this question would be (60-75)^2+(70-75)^2+(80-75)^2+(90-75)^2+(100-75)^2.

We used a Mann-Whitney U test (Wilcoxon Rank Sum test) for all analyses and treated differences as statistically significant if the two-sided p-value was less than 0.05. We performed additional tests using parametric t-tests to confirm that our results were not driven by the choice of statistical test.

Results

Primary analysis

Tables 1-4 summarizes comparisons across all conditions for both accuracy measures. Recall that more negative values indicate greater accuracy. For our primary accuracy measure (DV1), we found that median accuracy in the Reciprocal Scoring condition ($M_{Reciprocal}$ = -0.52, $SE_{Reciprocal}$ = 0.05) was indistinguishable from the Brier condition (M_{Brier} = -0.53, SE_{Brier} = 0.05), non-significant at an alpha level of 0.05 and a two-tailed test (p = 0.40). We also found that participants in the Reciprocal Scoring condition were more accurate ($M_{Reciprocal}$ = -0.52, $SE_{Reciprocal}$ = 0.05) than participants in the Control condition ($M_{Control}$ = -0.34, $SE_{Control}$ = 0.05), using a two-tailed test (p = 0.001). Appendix B shows question-by-question measures of accuracy and tests of statistical significance between group.

Table 1: DV1 Summary

Condition	N	Median	Mean	Trimmed Mean	SD	SE
C1: Control	433	-0.34	0.00	-0.27	1.00	0.05
C2: Brier	423	-0.53	-0.18	-0.44	0.94	0.05
C3: Superforecaster	17	-0.87	-0.83	-0.89	0.24	0.06
C4: Reciprocal	428	-0.52	-0.12	-0.39	1.01	0.05

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Table 2: DV2 Summary

Condition	N	Median	Mean	Trimmed Mean	SD	SE
C1: Control	433	-0.32	0.00	-0.27	1.00	0.05
C2: Brier	423	-0.61	-0.29	-0.54	0.91	0.04
C3: Superforecaster	17	-0.82	-0.62	-0.81	0.61	0.15
C4: Reciprocal	428	-0.51	-0.16	-0.42	1.00	0.05

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Table 3: DV1 Mann Whitney U Test (Wilcoxon Rank Sum Test) Summary

Group 1	Group 2	P value
C2: Brier	C1: Control	<0.001
C3: Superforecaster	C1: Control	<0.001
C3: Superforecaster	C2: Brier	<0.001
C4: Reciprocal	C1: Control	<0.01
C4: Reciprocal	C2: Brier	0.40
C4: Reciprocal	C3: Superforecaster	<0.001

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Table 4: DV2 Mann Whitney U Test (Wilcoxon Rank Sum Test) Summary

Group 1	Group 2	P value
C2: Brier	C1: Control	<0.001
C3: Superforecaster	C1: Control	<0.001
C3: Superforecaster	C2: Brier	0.06
C4: Reciprocal	C1: Control	<0.001
C4: Reciprocal	C2: Brier	0.03
C4: Reciprocal	C3: Superforecaster	0.02

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Figure 1 ranks forecasters in each condition by accuracy (Brier score) and then for each value of the X-axis (N), plots on the Y-axis the median accuracy of forecasters 1 through N. The rightmost points show the median accuracy across all forecasters in each condition (as in Table 4), but we can see in this figure that the gap in accuracy between Brier and Reciprocal conditions is negligible at all levels of accuracy. And both Brier and Reciprocal conditions significantly outperform the control group across the distribution of accuracy.

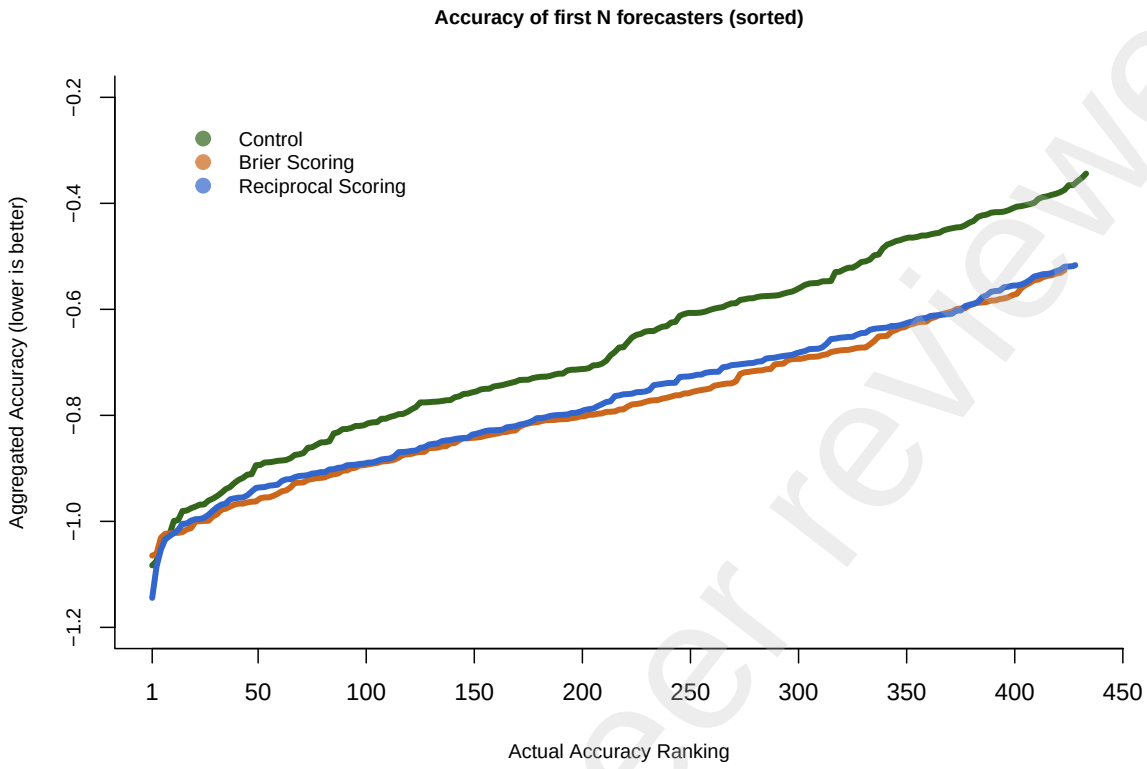


Figure 1: Forecasters are ranked by realized accuracy from most to least accurate (left-to-right) on the x-axis. On the y-axis we plot the median accuracy of forecasters 1 through N, by condition.

Comparing the secondary accuracy measure using the full distribution of forecasts for each question, we again find that participants in the Reciprocal Scoring condition ($M_{Reciprocal}$ = -0.51, $SE_{Reciprocal}$ = 0.05) and the Brier condition (M_{Brier} = -0.61, SE_{Brier} = 0.04) were more accurate than participants in the Control condition ($M_{Control}$ = -0.32, $SE_{Control}$ = 0.05), two-tailed $p < 0.0001$ for Brier-Control contrast and $p = 0.0002$ for Reciprocal-Control contrast. We also find that accuracy was significantly greater in the Superforecaster condition than in any other condition for the primary accuracy measure, better than both the Reciprocal Scoring and Control conditions for the secondary accuracy measure, and statistically indistinguishable from the Brier condition for the secondary accuracy measure.

Exploratory analysis

We now explore two plausible proxies for forecasting effort: survey response time and the

number of sources used. Using a Mann-Whitney U test (Wilcoxon Rank Sum) of distributional differences in response times in the Brier, Reciprocal, and Control conditions, we found that median response times in the Brier conditions (M_{Brier} = 41.40 , SE_{Brier} = 0.93) were slightly lower than in Reciprocal conditions ($M_{Reciprocal}$ = 43.80, $SE_{Reciprocal}$ = 1.16), nonsignificant on a two-tailed test, p = 0.48.

Second, we conducted a Mann-Whitney U test of the distributional differences in the number of information sources that participants indicated using per question, again comparing the Brier and Reciprocal Scoring conditions. We found that the mean number of sources used in the Reciprocal Scoring conditions ($M_{Reciprocal}$ = 1.35, $SE_{Reciprocal}$= 0.06) was almost the same as in the Brier conditions (M_{Brier} = 1.5 sources, SE_{Brier} = 0.07), a non-significant difference, two-tailed p = 0.72. Table 5-6 summarizes the comparisons across conditions for the measure of response times and the measure of the number of sources consulted.

Table 5: Summary of the Time Spent (In Minutes) on the Whole Survey

Condition	N	Median	Mean	Trimmed Mean	SD	SE
C1: Control	433	37.10	41.45	43.99	19.32	0.93
C2: Brier	423	41.45	46.01	48.76	19.10	0.93
C3: Superforecaster	17	134.52	293.97	327.14	460.39	111.66
C4: Reciprocal	428	43.82	48.08	51.21	24.05	1.16

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Table 6: Summary of the Number of Sources Consulted per Question

Condition	N	Median	Mean	Trimmed Mean	SD	SE
C1: Control	433	1.00	1.64	1.71	1.38	0.07
C2: Brier	423	1.50	1.77	1.86	1.45	0.07
C3: Superforecaster	17	2.00	2.01	2.12	0.94	0.23
C4: Reciprocal	428	1.35	1.74	1.83	1.21	0.06

Note: The trimmed mean is the mean of the accuracy of forecasters in each condition, after deleting the 10% least accurate forecasts from that condition.

Discussion

Reciprocal Scoring extends the analytic reach of conventional forecasting tournaments by enabling evaluations of forecasts of objectively unresolvable questions. The goal of Study 1 was to examine whether this expansion of range comes at a cost in rigor. We find that forecasts elicited by Reciprocal Scoring are no worse than those elicited by Brier scoring on our primary accuracy measure and only slightly worse using a secondary accuracy measure. For both measures, participants incentivized to do well on either Brier or Reciprocal Scoring do significantly better than those with no incentive to be accurate. Moreover, Reciprocal Scoring and Brier scoring elicit similar effort as measured by time spent on questions and sources used. Taken together, these results suggest that Reciprocal Scoring does indeed incentivize rigorous forecasting and can be used to estimate likely impacts of alternative policy options during ongoing crises such as the COVID-19 pandemic.

Study 2: COVID-19 policy evaluation tournament

Tournament setup

To put Reciprocal Scoring to work, we recruited two teams of a dozen forecasters, all of whom had excelled in previous forecasting exercises, like ACE, FOCUS, or public platforms (e.g., Good Judgment Open). Each participant forecasted two baseline questions: the total number of U.S.

deaths from COVID-19 in 2020 and 2021. Participants also forecasted total COVID-19 deaths conditional on each of eleven unique policies being implemented individually and immediately across the U.S, as well as two bundles of multiple policies being (hypothetically) implemented in conjunction – see Table 8 for an overview of all the policy hypotheticals. The aim was to infer the estimated effect of each individual policy from the difference in estimates between the baseline forecasts and each policy forecast.

Table 7: Counterfactual policies in the COVID-19 tournament

Policy	**Description**
Baseline*	Total U.S. deaths attributed to COVID-19 throughout the calendar year.
Federal Mask Mandate	Mandating mask usage in all public spaces (eg. on most sidewalks and while using public transportation)
Closing Schools*	Banning the in-person reopening of all K-12 public schools
Closing Universities	Banning the in-person reopening of all public colleges and universities
Ban Gatherings >50	Banning all non-essential gatherings of more than 50 people
Ban Gatherings >10	Banning all non-essential gatherings of more than 10 people
Stay-at-Home Order*	Requiring everyone except for essential workers to stay at home for three weeks, only leaving their homes for essential visits (i.e., those to doctors, hospitals, and grocery stores)
Restrict Indoor Dining	Closing all indoor dining at restaurants and bars
$20B Vaccine Prize*	Offering a 20 billion dollar prize to the first company that provides to the U.S. government enough vials of vaccine to vaccinate 50 million Americans. The vaccine must be FDA-approved and it must reduce the probability of being infected by Covid-19 by at least 50%
Mask Production and Distribution	Ordering the production of 10 billion disposable N95-equivalent masks, mailing the masks to all Americans, and distributing them widely for free. Assume that everyone in the U.S. is mailed at least ten masks by the end of August.
Federal Contact Tracing Program	Facilitating the hiring and training, by the U.S. Census Bureau, of 1 million people to contact trace Covid-19 cases
$100B for Testing	Earmarking an additional 100 billion dollars for Covid-19 testing, providing an additional $100 in reimbursement from the Federal government to testing laboratories for each FDA-approved positive and negative Covid-19 test completed in 2020.
Short Term Bundle*	Combination of mask mandate, stay-at-home order, and mask production and distribution.
Long term bundle*	Combination of closing schools, vaccine prize, and federal contact tracing program

Notes: For all policies, participants submitted forecasts of 2020 mortality. To ease the burdens, we only asked about 2021 for the subset of policies marked with an asterisk * in the table. In both cases, participants were told to assume the policies were in place for all of 2020, but not throughout 2021. Participants were also told to assume that the policies would be implemented individually; i.e., for each policy, all else would be held equal. We selected policies prominently featured in current policy debates and examined extensively using other methods (e.g., Brauner et al., 2020).

Participants provided their 5th, 25th, 50th, 75th, and 95th percentile estimates of total COVID-19 mortality. They were told that, in addition to their compensation for participation, they would be awarded an accuracy bonus determined by Reciprocal Scoring. Each forecaster would be paid a linear prize as a function of the average squared difference between that individual's forecasts and the median forecast from the other team. Drawing on published epidemiological models, clinical literature, news stories, and their own calculations, participants recorded their forecasts on a custom platform with message boards and tools for visualizing personal and median team forecasts. The tournament spanned two weeks, including four initial days of 'independent judgment mode,' during which time forecasters attempted the questions without accessing message boards or crowd forecasts. The analysis of this tournament was not pre-registered.

Results

Forecasters answered all questions, with an average of 4.11 forecasts on each question, and updating forecasts most often for the overall U.S. mortality in 2020. Given our interest in best-estimate forecasts after the two-week tournament, we focus on the final forecasts for each question. For each mortality baseline and policy question, we computed the median forecast for both of the independent teams on each forecasted quantity. To examine whether forecasts between teams were correlated, we compared median forecasts made by both teams across all policies and elicited percentiles. At the 50th percentile, the forecasted 2020 mortality reduction for the proposed policies had a Spearman rank-rank correlation of 0.83 ($p < 0.001$), while those for 2021 had a rank-rank correlation of 0.85 ($p < 0.001$), suggesting strong convergence of forecasts despite the complete independence and the absence of communication between teams. This convergence implies that Reciprocal Scoring with teams of historically accurate forecasters can lead to stable forecasts of the relative causal impact of policy options. Figure 2 highlights this convergence by comparing the ordering of expected 2020 mortality reduction, for each policy at the 50th percentile, as forecasted by Team A and Team B. Where lines intersect, there was disagreement between teams about the relative effectiveness.

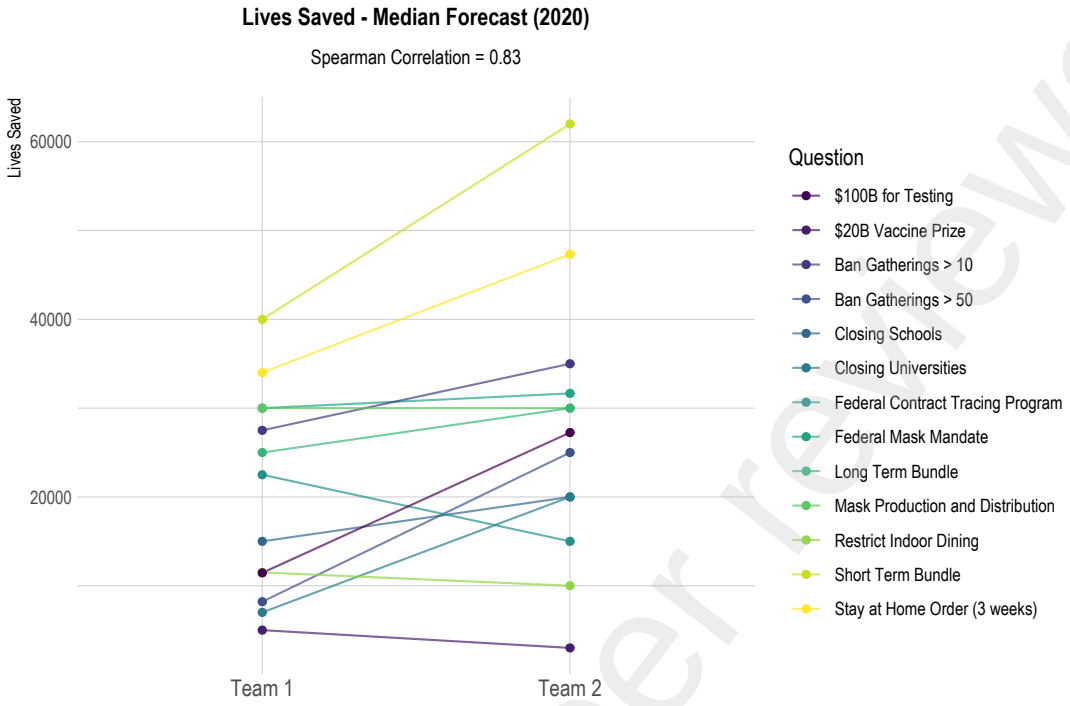


Figure 2: Comparison of median forecasted mortality reduction, calculated as the difference between the median policy-specific mortality total for each team and that team's baseline forecast for deaths in 2020. See Appendix C for similar graphs of the 25th and 75th percentiles, which had Spearman correlations of 0.60 and 0.74, respectively.

Given the between-team similarity – and drawing on the robust finding in the forecasting literature that averaging forecasts of independent groups tend to reduce noise and cancel opposing biases– we collapsed the two teams into a single group for the remaining analyses.

Figure 3 shows the distribution of final forecasts by each participant for each of the five percentiles for which we elicited 2020 and 2021 baseline predictions, and Appendix D contains the median forecast among all participants for each policy and percentile.

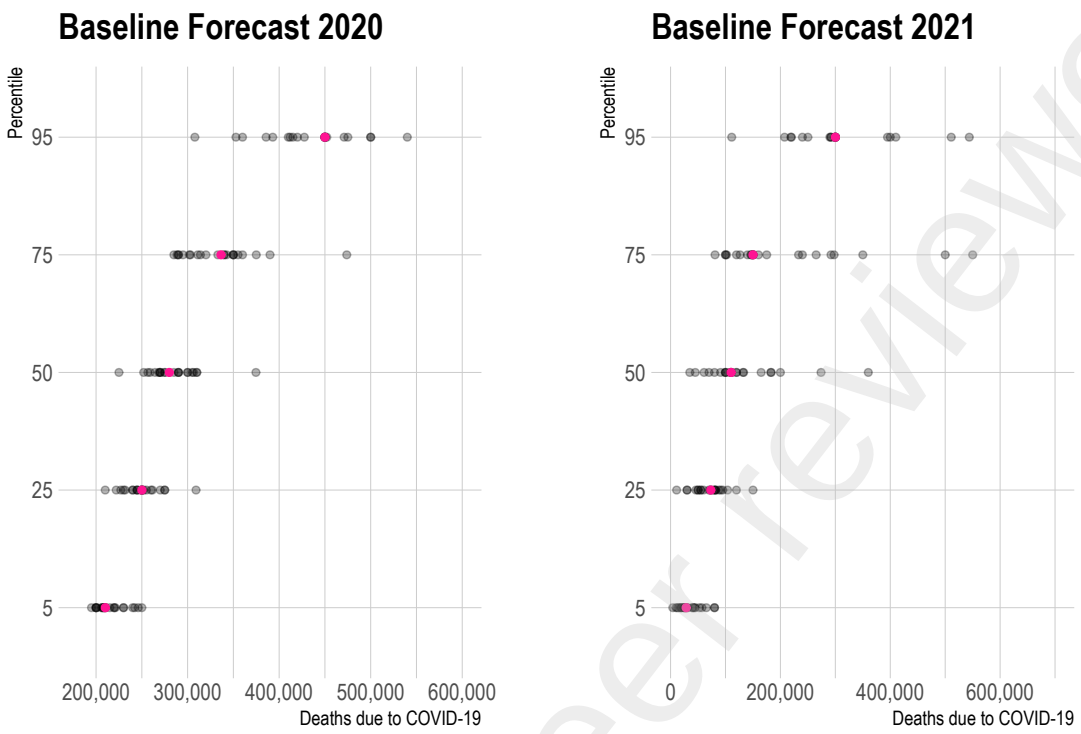


Figure 3: Final forecasts for the two baseline questions about overall COVID-19 mortality in the U.S. for 2020 and 2021. Black dots represent individual forecasts and red dots represent median forecasts for each percentile. These figures form a cumulative distribution function of forecasted COVID-19 deaths in 2020 and 2021.

As Figure 3 shows, forecasters anticipated fewer deaths in 2021 than in 2020 across all percentiles, with a median forecast of 280,000 for 2020 and 110,212 for 2021. Although the forecasts were recorded in early August, the baseline forecasts for 2020 are comparable to other projections as of late September (e.g., Gu 2020). Actual COVID-19 deaths in 2020 and 2021 were significantly higher than predicted, with the U.S. reporting just under 379,000 deaths in 2020 (close to the 75th percentile of forecasted COVID-19 deaths from our study) and just under 317,000 deaths through September 29th, 2021.[10]

By forecasting a range of policies across the same periods, our tournament enables a comparison of relative policy effectiveness across percentiles of the expected probability distribution.

[10]Source: CDC COVID Data Tracker, accessed on October 1st, 2021 at https://covid.cdc.gov/covid-data-tracker/#trends_dailydeaths.

Figure 4 shows forecasted 2020 mortality under each policy, for each of the five percentiles, including median forecasts and bootstrapped 95% confidence intervals.[11]

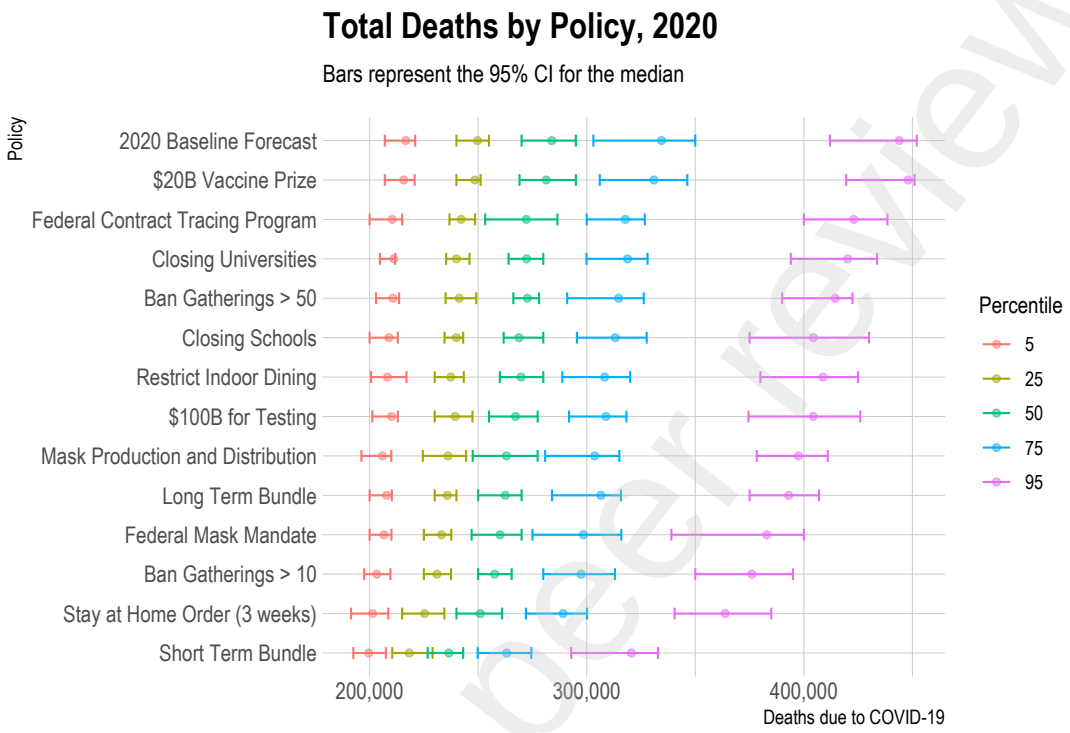


Figure 4: *Estimated COVID-19 mortality by policy, 2020.* Circles represent the median forecast for each percentile and bars represent the bootstrapped 95% CI for the median. The color scheme represents each of the five percentiles. Forecasts are in descending order according to the median forecast for the 50th (green) percentile.

We inferred the impact of each policy by subtracting the forecasted mortality for each policy from the baseline forecasts; see Appendix E for median forecasts of mortality reduction. Figure 5 displays the estimated impact of all policies on COVID-19 mortality in 2020 for the 50th percentile.

[11] Our final sample consisted of 22 highly experienced and historically accurate forecasters. We bootstrap (N=10,000) the confidence intervals here by drawing 22 forecasters with replacement from our sample of forecasters, for each percentile forecast, for each policy, and calculating the median forecast for that percentile and that policy. The confidence intervals represent the 95th percentile of the empirical distribution of bootstrapped median forecasts.

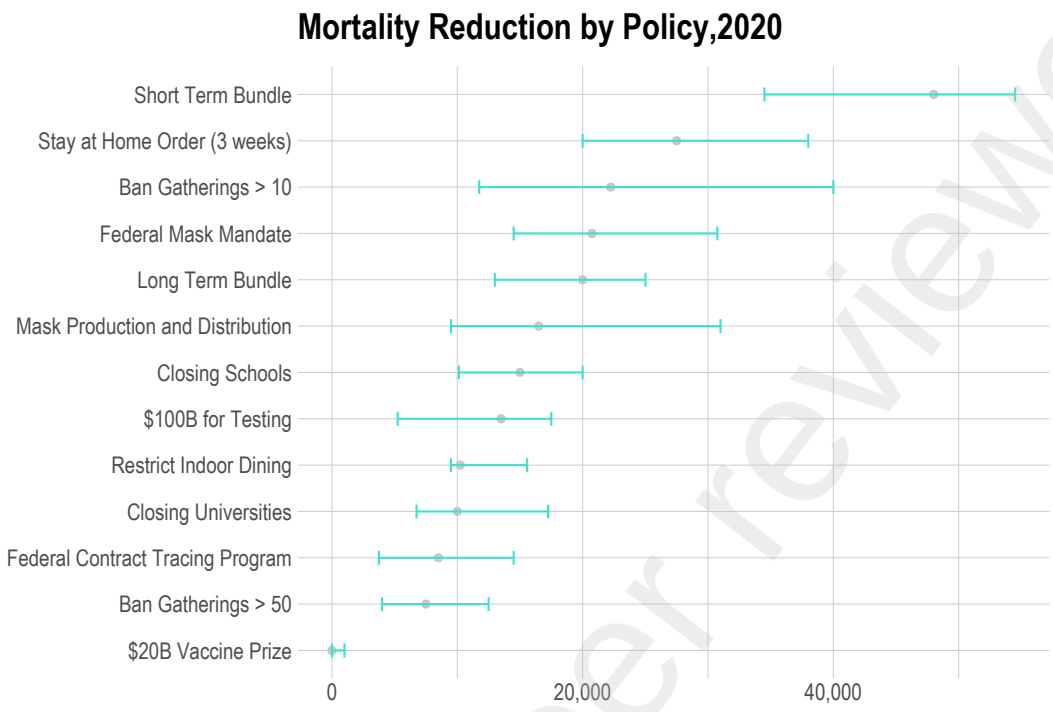


Figure 5: Estimated COVID-19 mortality reduction at the 50th-percentile by policy, 2020. Circles represent the difference between the median baseline forecast and median policy forecast for each policy at the 50th percentile. Bars represent 95% confidence intervals, which we computed by subtracting each of the upper and lower bounds of the bootstrapped policy estimates from the upper and lower bounds of the bootstrapped baseline estimate.

Reciprocal Scoring forecasts suggest that a policy bundle including a federal mask mandate, mask production, and a three-week stay-at-home order would have reduced mortality the most: by over 40,000 deaths at the 50th percentile. Longer-term policies, such as a large prize for successful vaccine development, had smaller expected effects on 2020 COVID-19 mortality, perhaps because forecasters saw vaccines as unlikely to be a major driver of short-term mortality or because they did not see monetary constraints as a serious bottleneck for vaccines then furthest along the development pipeline. Importantly, these estimates are of the *marginal* effects of each policy: a federal ban on large gatherings, for example, may be estimated to have a limited effect in the current world, where such gatherings are already rare because of local bans or endogenous changes in behavior.

Figure 6 visualizes the estimated mortality reduction across the entire distribution of COVID-

19 deaths in 2020. We compare the median forecast for each of our elicited percentiles of the mortality distribution (5%, 25%, 50%, 75%, 95% in the baseline world with no sudden policy changes) to the hypothetical world with a sudden policy change. We then fit a self-starting logistic distribution[12] to the median forecast of deaths at each percentile, for each policy.[13] And we separately plot the forecasted distribution of deaths for each policy and at baseline (with no sharp policy changes). For example, in Panel A we show the forecasts suggesting that a federal mask mandate would reduce mortality by 30,000 at the median, with relatively lower effects on COVID-19 mortality at the 5th percentile. By contrast, a vaccine prize (Panel C) is forecasted to have minimal effects on short-term 2020 COVID-19 mortality across the entire distribution.

[12] A logistic distribution was selected from a family of shortlisted choices (polynomial, beta distribution, and logistic) because it best satisfied the endpoint conditions (100th percentile forecast = 1, 0th percentile forecast = 0).

[13] This imputation process imposes a parametric fit on our nonparametric-elicited forecasts, but as Figure 6 shows, the smoothness of the parametric model only accentuates the policy effects implied by the nonparametric-elicited probabilities. The one weakness to this fit is in imputing the forecasting distribution at the endpoints (forecasting $<10\%$ or $>90\%$ values).

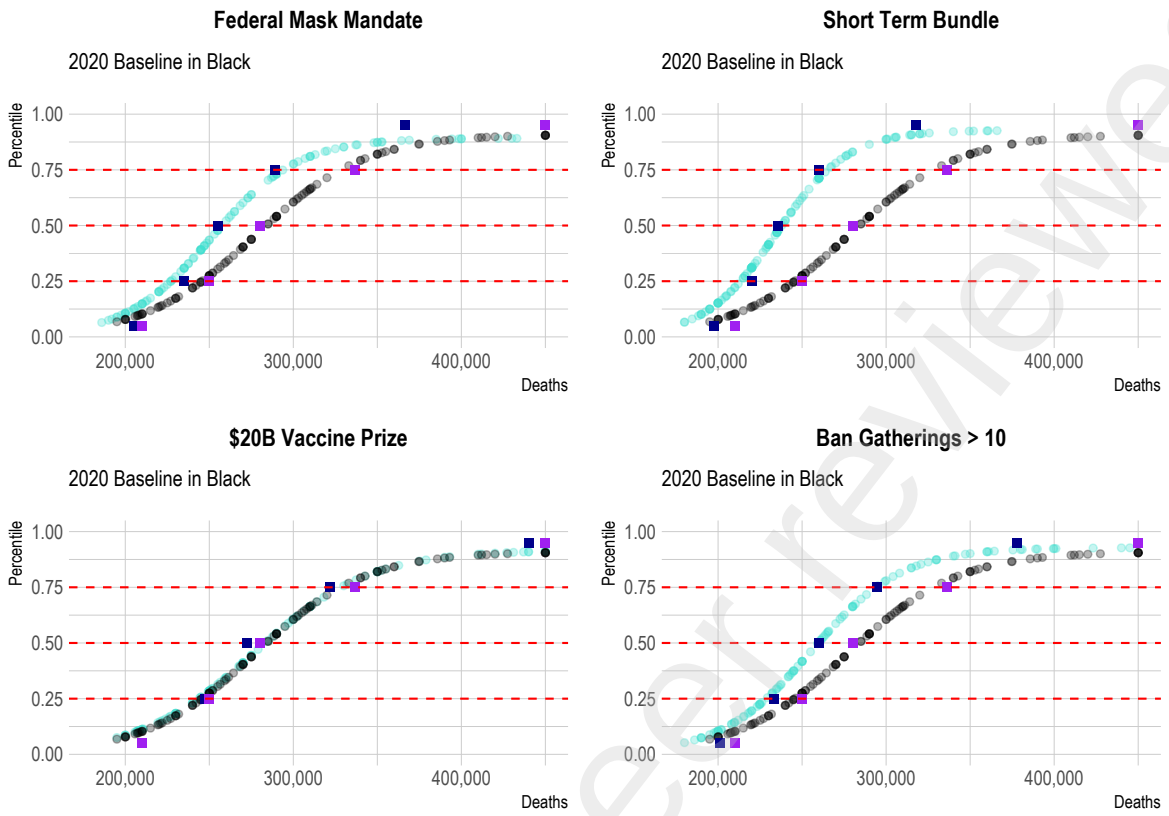


Figure 6: Median forecasts for each policy are represented by purple (baseline) and blue (policy counterfactual) squares. The turquoise line represents a logistic fit to the blue squares, and the black line represents a logistic fit to the purple squares. The plot shows the cumulative distribution function of forecasted COVID-19 mortality with policy implementation relative to baseline COVID-19 mortality. Appendix F includes the remaining policies.

The merits of Reciprocal Scoring for policy evaluation

Beyond its implications for debates about optimal COVID-19 policy responses in the summer of 2020, our tournament provided the first test of the policy-evaluation utility of Reciprocal Scoring. The biggest benefit of Reciprocal Scoring is that it permits the prospective evaluation of potential policies without the incentive problems that arise when forecasting unresolvable questions. However, the methodology offers several other advantages over many existing approaches to evaluating policies. First, it allows for a direct head-to-head comparison of policies that otherwise might be elusive if interventions are evaluated individually using different metrics. Moreover, by including bundles of policies, researchers can flexibly explore interactions among interventions.

For example, it can be unclear whether combinations of policies will curb exponential spread better, or whether policy bundles will be less than the sum of their parts; e.g., because the marginal benefit of reducing social contact could be lower if high usage of protective masks reduces transmission risk during social contact. In our tournament, the short-term bundle was forecasted to reduce mortality by less than the sum of the independent forecasts for the constituent policies, revealing an assumption of a negative interaction among policies.

A potential concern is that these benefits are gained at the expense of the quality of the obtained estimates. The results of Study 2 should assuage this concern, at least somewhat. The tournament produced rankings of forecasts that intuitively seem both coherent and plausible. Moreover, the between-team correlation of forecasts shows that Reciprocal Scoring can avoid yielding bimodal forecast distributions with ambiguous interpretations. In addition to this indirect evidence, we asked participants directly about their behavior under the Reciprocal Scoring paradigm, using an anonymous survey after the tournament. We wanted to examine whether forecasters had tried to game Reciprocal Scoring as discussed earlier, or whether they had simply reported their true beliefs, as our game-theory framework predicted. Among the 21 tournament participants, 19 completed the full survey. Of those, 12 (63%) responded "I always reported my exact true beliefs" while seven (37%) responded "I sometimes reported something other than my exact true belief." On average, these seven respondents indicated that there was only a minimal 15% difference between their true beliefs and the forecasts they reported, and the reasons for this divergence varied – see Appendix G for an overview.

Because Reciprocal Scoring is designed for questions that cannot be formally resolved, it will rarely be possible to directly evaluate the quality of forecasts elicited under the method. Nonetheless, subsequent empirical analyses can offer instructive sanity checks. Consider the example of the COVID-19 policy forecasts generated in Study 2. Given that the forecasts focused on specific interventions at a specific point in time, we should be wary of comparisons with subsequent analyses of similar – but ultimately distinct – interventions. However, because the general types of interventions examined in Study 2 have been subjected to considerable subsequent scrutiny, we can

draw on interdisciplinary analyses to gauge the congruence between Reciprocal Scoring forecasts and ex-post intervention evaluations. While comparing effect sizes across studies is complicated by the fact that different analyses define interventions differently (e.g., mask mandates versus mask recommendations) and quantify effects in different terms (e.g., reductions in new COVID-19 cases versus reductions in the reproduction rate R), it is still possible to make broad inferences about how well the Study 2 forecasts align with retrospective analyses that have the benefit of hindsight.

For example, in a study of 114 subnational areas in Europe, Sharma et al. (2021) found that restrictions on large gatherings (above 30 people) were among the least effective of numerous non-pharmaceutical interventions studied using a hierarchical Bayesian model focusing on the reproduction rate, a strong correlate of mortality. This finding corresponds with Study 2, which ranked the banning of gatherings above 50 as the second least effective policy, behind only the vaccine prize. Sharma et al. also found evidence to suggest that stricter mask-wearing policies were likely to be more effective than the closure of educational institutions, consistent with the Study 2 ranking.

Similarly, regression analyses Liu et al. and by Leffler et al. (2020) compared several of the interventions from Study 2 – including school closures, restrictions on gatherings, and stay-at-home orders – and found that federal contract tracing programs were among the interventions with the relatively weakest evidence for an association with reduced mortality. This finding is also consistent with Study 2, in which a federal contact tracing program was forecasted to be the fourth-least effective among 13 proposed policies.

For some interventions, however, published analyses may differ from Study 2 forecasts. For example, Study 2 ranked a three-week stay-at-home-order as the single intervention with the highest expected effect on 2020 mortality, while the evidence for the effectiveness of stay-at-home orders relative to other interventions has been weak in various studies, including by Liu, Morgenstern et al., Leffler et al., and Stokes et al. Such incongruence may be partly explained by the fact that stay-at-home orders are defined and operationalized in different ways. For example, the Study 2 version of the policy was defined in considerably stricter terms than many of the recommendations

or requirements that were in fact implemented in the United States; and in the countries where stay-at-home orders have been a more integral part of successful public health responses, such as China, Israel, and Australia, the policies have typically been implemented in the strict manner similar to the proposed policy in Study 2.

Discussion

Study 1 shows that Reciprocal Scoring can elicit estimates as accurate as the methods that are conventionally used in human-judgment forecasting. However, the real value-proposition of Reciprocal Scoring is its flexible and policy-relevant applications. As Study 2 showed, Reciprocal Scoring can produce timely evaluations of competing decision-options, such as public health policies. So, the method can give policy-makers faster access to better probability-judgments of alternative options than they could have gotten by waiting for the resolution of unconditional questions in first-generation forecasting tournaments and prediction markets. These options can be designed to be both interpretable and actionable, allowing for immediate policy-relevance following the submission of forecasts. Moreover, the method is not limited to evaluating any single outcome; unlike most other analytical tools, Reciprocal Scoring can be scaled to evaluate multiple effects of a single policy without any fundamental methodological adaptations. This makes Reciprocal Scoring an attractive method for public as well as private organizations facing choices among options with complex cost-benefit properties that are challenging – but not impossible–to comprehend ex-ante. Most importantly, Reciprocal Scoring is transportable. It can be used to incentivize effortful and accurate forecasting in a wide range of domains where objective resolution criteria are elusive. Beyond the evaluation of potential policies exemplified in Study 2, one such domain is forecasting long-run outcomes where question-resolution may not arrive soon enough to be useful for incentivizing forecasters. Accordingly, Reciprocal Scoring is uniquely suited for efforts to understand long-run trends in macroeconomics, geopolitics, technological development, and existential risks, as well as a host of other policy-relevant domains. Indeed, the method can even be applied to counterfactual claims about how history would have unfolded if policy-makers had made different

choices, allowing for the resolution of longstanding – and previously intractable – disputes.

We view Reciprocal Scoring as a novel methodology, not an isolated technique with a singular application. Studies 1 and 2 operationalized the methodology in a very particular way: a two-panel design, relying on "superforecasters" as the aspirational cognitive ideal. However, Reciprocal Scoring can be adapted to multi-sided disputes. For instance, Scoblic et al. 2021 identified four distinct schools of geopolitical thought on China and argued that the best forecasters would be those skilled at predicting the predictions that each school of thought would generate. This approach can be viewed as a variant of Reciprocal Scoring, with a key modification. The political leanings of superforecasters are supposed to be unknowable (or at least extremely difficult to infer without in-depth analysis of the issues) whereas the political leanings of the panels on China are known to fall within a certain plausibility range and the challenge lies in accurately translating qualitative verbal arguments into probability estimates.

The Scoblic et al approach and the approach in Study 2 rest on different assumptions about efficient pathways to truth discovery. Scoblic et al see value in distinguishing key schools of thought, in tacitly supposing that each school has captured a fraction of the truth, and in encouraging forecasters to master each school before trying to integrate perspectives into their own forecasts. The approach taken here invests more trust in the forecasters trying to forecast the elite panel—and assumes that in doing their due diligence, they will survey the range of plausible truth-containing perspectives in their own fashion and need no help in discovering what those perspectives are. Our approach is thus a hands-off application of Reciprocal Scoring, whereas the Scoblic et al approach is a guided variant of our use of Reciprocal Scoring. The "hands-off" approach has the advantage that it is more efficient when forecasters are highly competent but it is harder to trace sources of error when forecasters make mistakes. The guided approach has strengths and weaknesses: it should help to structure the task for less sophisticated forecasters and facilitate tracing of mistakes, but it imposes unnecessary bureaucratic burdens on highly competent forecasters.

In sum, the two studies reported here constitute a preliminary proof-of-concept of the epistemic and practical value of Reciprocal Scoring, a flexible methodology whose core advantage is

its extensionality: we can stretch it into domains where objective metrics of judgmental accuracy are either impossible or impractical but intersubjective metrics are feasible. Study 2 focused on an obvious high-value application: estimating the effects of immediately implementing each of an array of policy options under active debate. But the range of possible applications of Reciprocal Scoring is far wider and encompasses all tasks that have high expected value but, given the state of early 21st century technology, rely on subjective human judgment. For instance, why not task Reciprocal Scoring panels with generating creative forecasting questions for tournaments and markets, or with judging the appropriate lessons to extract from history in policy post-mortems?

References

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

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, *12*(2), 90–103.

Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenćiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y., & Kulveit, J. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science*, *371*(6531), Article 802. https://doi.org/10.1126/science.abd9338

Chang, W., Chen, E., Mellers, B., & Tetlock, P. E. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526.

Court, D., Gillen, B., McKenzie, J., & Plott, C. R. (2018). Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice Prophecy and Guess of Guesses. *Economic Theory*, *65*, 25–54.

Cvitanić, J., Prelec, D., Riley, B., & Tereick, B. (2019). Honesty via choice-matching. *American Economic Review: Insights*, *1*(**??**), 179–192. https://doi.org/10.1257/AERI.20180227

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Perez-Guzman, P. N., Schmit, N., Cilloni, L., Ainslie, K. E. C., Baguelin, M., Boonyasiri, A., Boyd, O., Cattarino, L., ...Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*(7820), 257–261. https://doi.org/10.1038/s41586-020-2405-7

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PLoS ONE*, *12*(5), e0177385. https://doi.org/10.1371/JOURNAL.PONE.0177385

Grice, H. P. (1967). Logic and conversation. In P. Grice (Ed.), *Studies in the way of words* (pp. 41–58). Harvard University Press.

Gu, Y. (2021). *COVID-19 projections using machine learning*. https://covid19-projections.com

Institute for Health Metrics and Evaluation (IHME). (2021). *COVID-19 Projections*. https://covid19.healthdata.org/projections

Ioannidis, J. P. A., Cripps, S., & Tanner, M. A. (2020). Forecasting for COVID-19 has failed. *International Journal of Forecasting*. Advance online publication. https://doi.org/10.1016/J.IJFORECAST.2020.08.004

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (2018). Prospect theory: An analysis of decision under risk. In L. C. MacLean & W. T. Ziemba (Eds.), *Handbook of the fundamentals of financial decision making* (pp. 99–127). World Scientific.

Keynes, J. M. (1936). *The general theory of employment interest and money*. Palgrave Macmillan.

Leffler, C. T., Ing, E., Lykins, J. D., Hogan, M. C., McKeown, C. A., & Grzybowski, A. (2020). Association of country-wide coronavirus mortality with demographics, testing, lockdowns, and public wearing of masks. *American Journal of Tropical Medicine and Hygiene*, *103*(6), 2400–2411. https://doi.org/10.4269/AJTMH.20-1015

Liu, Y., Wang, J., & Chen, Y. (2020). *Surrogate scoring rules* [Paper presentation]. Proceedings of the 21st ACM Conference on Economics and Computation, Virtual Event, Hungary. https://doi.org/10.1145/3391403.3399488

McCoy, J., & Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint,* arXiv:1703.04778. http://arxiv.org/abs/1703.04778

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Emlen Metz, S., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/XAP0000040

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832. https://doi.org/10.1016/J.IJFORECAST.2017.04.002

Metaculus. (2020, June 2). *A preliminary look at metaculus and expert forecasts*. https://www.metaculus.com/news/2020/06/02/LRT/

Myatt, D. P., & Wallace, C. (2012). Endogenous information acquisition in coordination games. *Review of Economic Studies*, *79*(1), 340–374. https://doi.org/10.1093/RESTUD/RDR018

Prelec, D. (2004). A Bayesian truth Serum for subjective data. *Science*, *306*(5695), 462–466. https://doi.org/10.1126/SCIENCE.1102081

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/NATURE21054

Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J.B., Monrad, J.T., Altman, G., Dhaliwal, G. and Finnveden, L. (2021). Understanding the effectiveness of government interventions in Europe's second wave of COVID-19. *medRxiv,* 2021.03.25.21254330. https://doi.org/10.1101/2021.03.25.21254330

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

Tetlock, P. E. (2017). *Expert political judgment* (New ed.). Princeton University Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction* (Reprint ed.). Broadway Books.

Waggoner, B., Chen, Y. (2014). Output agreement mechanisms and common knowledge. *Second AAAI Conference on Human Computation and Crowdsourcing*.

Walker, P. G., Whittaker, C., Watson, O., Baguelin, M., Ainslie, K. E. C., Bhatia, S., Bhatt, S., Boonyasiri, A., Boyd, O., Cattarino, L., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Donnelly, C. A., Dorigatti, I., Elsland, S. v., FitzJohn, R., Flaxman, S., Fu, H., Gaythorpe, K., Geidelberg, L., Grassly, N., Green, W., Hamlet, A., Hauck, K., Haw, D., Hayes, S., Hinsley, W., Imai, N., Jorgensen, D., Knock, E., Laydon, D., Mishra, S., Nedjati-Gilani, G., Okell, L. C., Riley, S., Thompson, H., Unwin, J., Verity, R., Vollmer, M., Walters, C., Wang, H. W., Wang, Y., Winskill, P., Xi, X., Ferguson, N. M., & Ghani, A. C. (2020). *The global impact of covid-19 and strategies for mitigation and suppression*. Imperial College London. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Global-Impact-26-03-2020.pdf

Witkowski, J., Atanasov, P., Ungar, L., & Krause, A. (2017). *Proper proxy scoring rules* [Paper presentation]. Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA.

Appendix A: Methodological details for Study 1

We aimed to recruit 1,500 participants through Prolific. A total of 1,479 participants ultimately completed the survey, each of whom was paid at a rate of $10.00 per hour (yielding an expected payment of $5 for this survey, which was estimated to take 30 minutes), with the possibility of earning an additional bonus of $25. 195 participants were excluded based on the pre-registered exclusion criteria, such as failing various comprehension and attention checks (N = 67) or completing the survey in less than 15 minutes (N = 56). 72 participants in the nonparametric conditions were excluded from analysis of accuracy because we could not fit a model to their forecast: they were either non-monotonic or identical across increasing percentiles on at least two questions. The final sample size from Prolific is 1,284. Additionally, we recruited 17 highly accurate forecasters – "Superforecasters" identified during the IARPA-funded ACE forecasting tournament – to complete the survey. The tables below show exclusion criteria and participant demographics.

Table A1: Exclusion criteria

Criterion	**Exclusions**	**Included in**
Completing survey in $<$15 minutes	56	Excluded from main pre-registered analysis. Reincluded for exploratory robustness analysis.
Failing either comprehension or attention checks	67	Excluded from main pre-registered analysis. Reincluded for exploratory robustness analysis.
Violating conditions for non-parametric elicitation on at least two questions	72	Excluded from main pre-registered analysis. Reincluded for exploratory robustness analysis.

Note: Some participants failed two or more criteria and were thus counted double in the table.

Table A2: Participant age

Age	Frequency	Percentage
$18-25$	255	19.86
$26-35$	459	35.75
$36-45$	275	21.42
$46-55$	164	12.77
>55	131	10.20
Total	1284	100

Note: Since we didn't ask superforecasters any demographic questions, the table doesn't include superforecasters' data.

Table A3: Participant gender

Gender	Frequency	Percentage
Male	524	40.81
Female	737	57.40
Non-binary	15	1.17
Undisclosed	8	0.62
Total	1,284	100

Note: Since we didn't ask superforecasters any demographic questions, the table doesn't include superforecasters' data.

Table A4: Participant race and ethnicity

Race/ethnicity	**Frequency**	**Percentage**
American Indian or Alaska Native	3	0.23
Black or African American	129	10.05
Asian	120	9.35
Hawaiian or Pacific Islander	1	0.08
Hispanic or Latino/Latina/Latinx	6	0.47
Middle Eastern or North African	4	0.31
White	935	72.82
Multiple racial	32	2.49
Prefer not to say	19	1.48
Other	35	2.73
Total	1,284	100

Note: Since we do not ask superforecasters any demographic questions, the table doesn't include superforecasters' data.

Participants took the survey two weeks before the date in question. Participants were provided one link to a resource with historical data for each question and asked to spend five minutes answering each. Table A_4 below shows the exact wording for each question.

Table A5: Forecasting questions, resolved on June 30th relating to a range of topics

Question	Content
New York COVID Cases	Please spend up to five minutes answering the following question: How many total cases of COVID-19 (confirmed and probable) will be reported in the city of New York on June 30th?
U.S. COVID Vaccine	Please spend up to five minutes answering the following question: What is the total number of people in the United States who will have received at least one dose of the COVID-19 vaccine by June 30th?
India COVID Vaccine	Please spend up to five minutes answering the following question: What is the total number of people in India who will have received at least one dose of the COVID-19 vaccine by June 30th?
India COVID Cases	Please spend up to five minutes answering the following question: How many new cases of coronavirus disease 2019 (COVID-19) will be confirmed in India on June 30th?
India COVID Deaths	Please spend up to five minutes answering the following question: How many new deaths from coronavirus disease 2019 (COVID-19) will be confirmed in India on June 30th?
Oil Price	Please spend up to five minutes answering the following question: What will the daily price of Brent crude oil be on June 30th (in US dollars)?
Stock Price	Please spend up to five minutes answering the following question: What will the end-of-day value of the S&P 500 Index be on June 30th?
New York Temperature	Please spend up to five minutes answering the following question: What will the maximum daily temperature (Fahrenheit) be in Central Park, New York City on June 30th?
Sydney Temperature	Please spend up to five minutes answering the following question: What will the maximum daily temperature (Fahrenheit) be in Sydney, Australia on June 30th?
Earthquakes	Please spend up to five minutes answering the following question: On June 30th, how many earthquakes of magnitude 1.5 or greater will there be in the world?

Appendix B: Study 1 Median forecast by question

COVID-19 new reported cases in NYC on June 30th

Condition	Control	Brier	Reciprocal	Superforecaster
5%	106	122	122	87
25%	178	167.5	179	126
50%	243	217	243	165
75%	365	274.5	313	226
100%	491	347.5	413	270

Total number of people received at least one dose of COVID-19 vaccine in U.S. by June 30th

Condition	Control	Brier	Reciprocal	Superforecaster
5%	172,652,174	175,043,478	175,043,478	175,282,609
25%	181,260,870	181,753,419	183,652,174	180,063,977
50%	190,347,826	188,913,043	190,826,087	184,608,696
75%	198,956,522	196,558,471	198,559,322	188,913,044
100%	209,956,522	203,815,994	209,765,250	192,739,130

Total number of people received at least one dose of COVID-19 vaccine in India by June 30th

Condition	Control	Brier	Reciprocal	Superforecaster
5%	181,884,058	195,652,196	195,652,174	210,144,928
25%	202,173,913	209,420,290	210,869,565	224,896,486
50%	221,014,493	225,000,000	231,159,420	246,517,232
75%	243,478,261	240,579,710	250,783,699	263,768,116
100%	263,768,116	257,142,857	275,362,319	284,420,290

COVID-19 new reported cases in India on June 30th

Condition	Control	Brier	Reciprocal	Superforecaster
5%	56,667	73,913	65,217	28,391
25%	91,304	95,406	100,000	50,725
50%	117,391	118,030	121,739	75,362
75%	144,928	131,754	149,275	92,754
100%	181,159	157,971	197,101	140,580

COVID-19 new reported deaths in India on June 30th

Condition	Control	Brier	Reciprocal	Superforecaster
5%	1,275	1,304	1,130	609
25%	2,000	1,857	1,855	1,014
50%	2,667	2,406	2,464	1,565
75%	3,147	2,870	2,929	2,058
100%	3,986	3,193	3,420	2,725

Oil price

Condition	Control	Brier	Reciprocal	Superforecaster
5%	59	60	60	61
25%	65	67	65	67
50%	71	72	71	72
75%	78	77	77	76
100%	86	83	83	81

S&P 500 value

Condition	Control	Brier	Reciprocal	Superforecaster
5%	3,764	3,962	3,870	3,906
25%	4,017	4,099	4,056	4,099
50%	4,200	4,204	4,210	4,257
75%	4,352	4,311	4,332	4,393
100%	4,586	4,500	4,514	4,565

Temperature in NYC (degrees Fahrenheit)

Condition	Control	Brier	Reciprocal	Superforecaster
5%	60	63	63	68
25%	67	70	70	73
50%	75	77	76	82
75%	82	83	84	89
100%	91	90	90	95

Temperature in Sydney (degrees Fahrenheit)

Condition	Control	Brier	Reciprocal	Superforecaster
5%	50	53	50	55
25%	55	58	56	58
50%	62	63	62	63
75%	68	68	68	67
100%	74	72	72	70

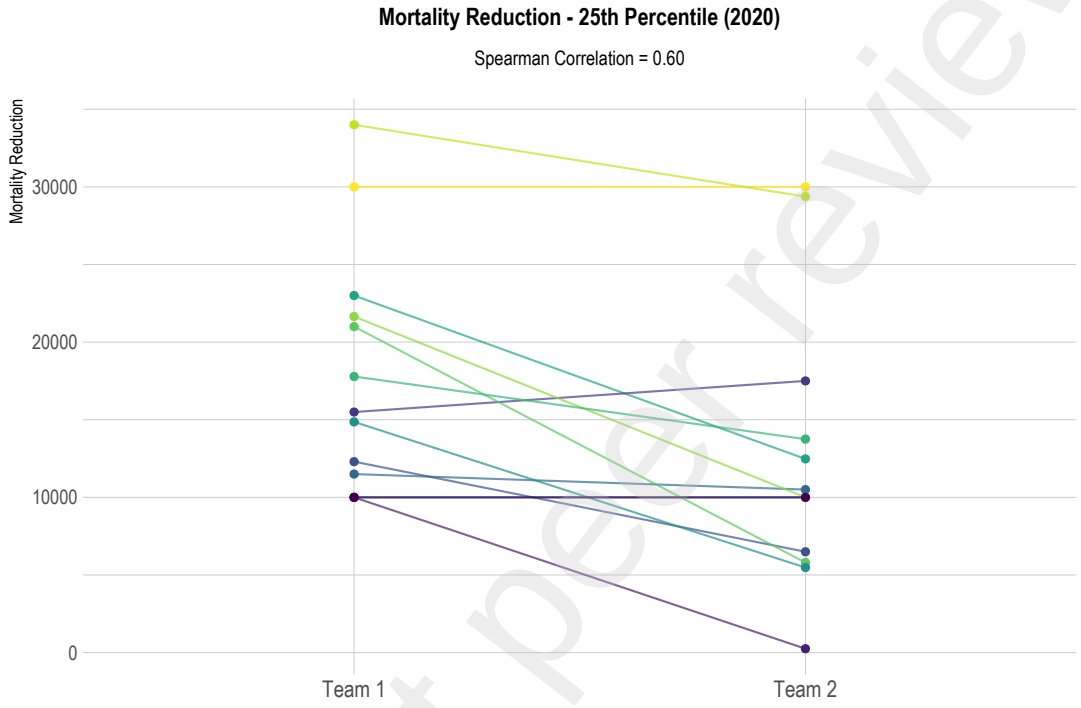
Number of earthquakes of magnitude 1.5 or greater in the world on Jun 30th

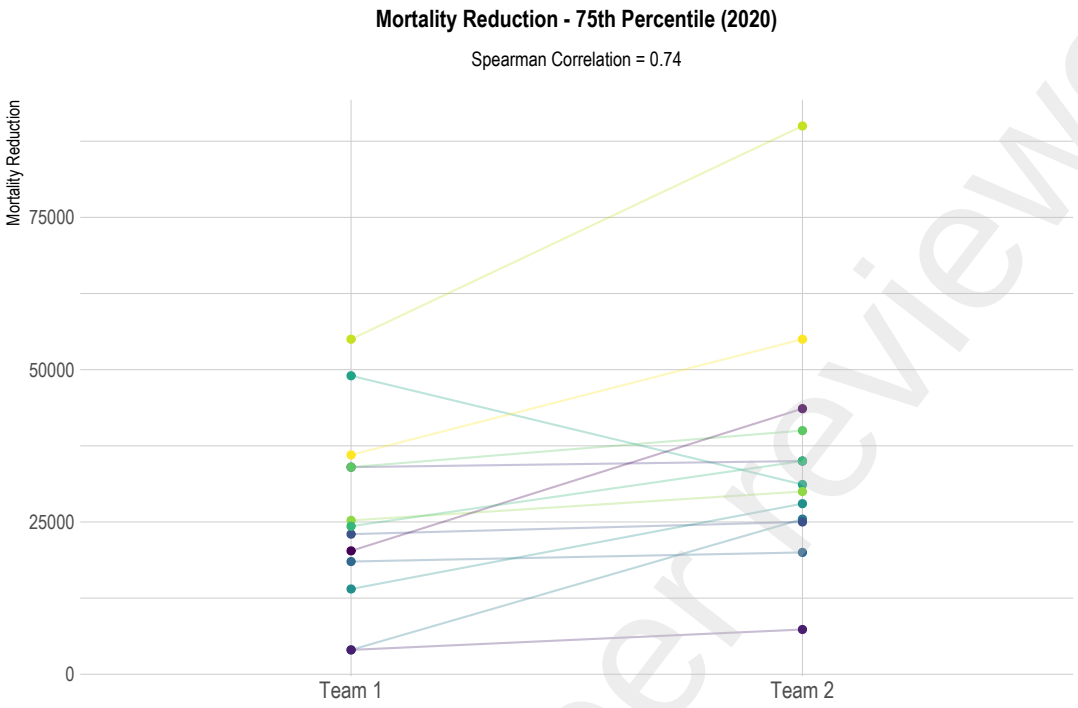
Condition	Control	Brier	Reciprocal	Superforecaster
5%	75	95	86	91
25%	100	113	109	139
50%	126	130	133	180
75%	153	152	160	247
100%	180	174	180	300

DV1 Statistics

	Median accuracy in Control group	Median accuracy in Brier group	Median accuracy in RS group	Median super forecaster accuracy	P-value from Brier-vs-Control	P-value from RS-vs-Control	P-value from RS-vs-Brier
New York COVID Cases	-0.31	-0.32	-0.32	-0.32	0.42	0.88	0.42
U.S. COVID Vaccine	-0.33	-0.34	-0.46	-0.36	0.14	0.55	0.06
India COVID Vaccine	-0.23	-0.26	-0.33	-0.99	0.33	0.02	0.10
India COVID Cases	-0.37	-0.37	-0.40	-0.41	1.00	1.00	1.00
India COVID Deaths	-0.33	-0.35	-0.35	-0.39	0.01	0.35	0.35
Oil Price	-0.36	-0.38	-0.38	-0.42	0.11	0.25	0.40
Stock Price	-0.31	-0.31	-0.31	-0.32	0.02	0.02	0.77
New York Temperature	-0.34	-0.41	-0.37	-0.55	0.01	0.05	0.45
Sydney Temperature	-0.35	-0.38	-0.39	-0.42	0.17	0.17	0.99
Earthquakes	-0.38	-0.46	-0.46	-0.71	0.04	0.06	0.86

Appendix C: Between-team comparisons of forecasts at the 25th and 75th percentile





Appendix D: Overall U.S. COVID-19 mortality by policy options

2020

Policy	Percentile				
	5th	25th	50th	75th	95th
2020 Baseline Forecast	210,157	250,000	285,000	333,000	450,000
Ban Gatherings > 50	209,821	239,250	273,000	314,000	411,000
Closing Universities	209,694	240,000	270,000	318,000	410,000
$20B Vaccine Prize	210,000	249,750	270,000	314,000	440,000
Closing Schools	206,500	239,500	266,800	310,000	398,700
Federal Contract Tracing Program	210,000	240,000	266,700	307,000	415,000
Restrict Indoor Dining	209,678	240,000	266,254	310,000	416,000
$100B for Testing	209,000	240,000	265,000	300,000	400,000
Ban Gatherings > 10	200,000	220,000	261,100	295,000	375,450
Mask Production and Distribution	200,000	234,000	260,000	306,200	400,000
Federal Mask Mandate	205,000	235,000	255,000	288,275	364,400
Long Term Bundle	207,000	235,000	255,000	303,750	394,000
Stay-at-Home Order (3 weeks)	200,000	232,500	250,000	285,000	360,000
Short Term Bundle	199,757	220,000	236,188	260,000	315,000

Note: Values reflect the median forecast among all participants for each policy and for each elicited percentile. Policies are ordered in descending order according to the 50th percentile.

2021

Policy	**Percentile**				
	5th	**25th**	**50th**	**75th**	**95th**
2021 Baseline Forecast	28,410	73,308	110,212	150,000	300,000
$20B Vaccine Prize	28,987	63,127	104,750	140,000	265,000
Stay-at-Home Order (3 weeks)	25,000	63,750	100,000	143,000	285,000
Closing Schools	28,410	62,500	97,375	133,000	250,000
Long Term Bundle	23,450	52,500	85,500	116,000	200,500
Short Term Bundle	15,000	40,000	65,627	90,000	143,000

Note: Values reflect the median forecast among all participants for each policy and for each elicited percentile. Policies are ordered in descending order according to the 50th percentile.

Appendix E: Overall U.S. COVID-19 mortality reduction by policy options

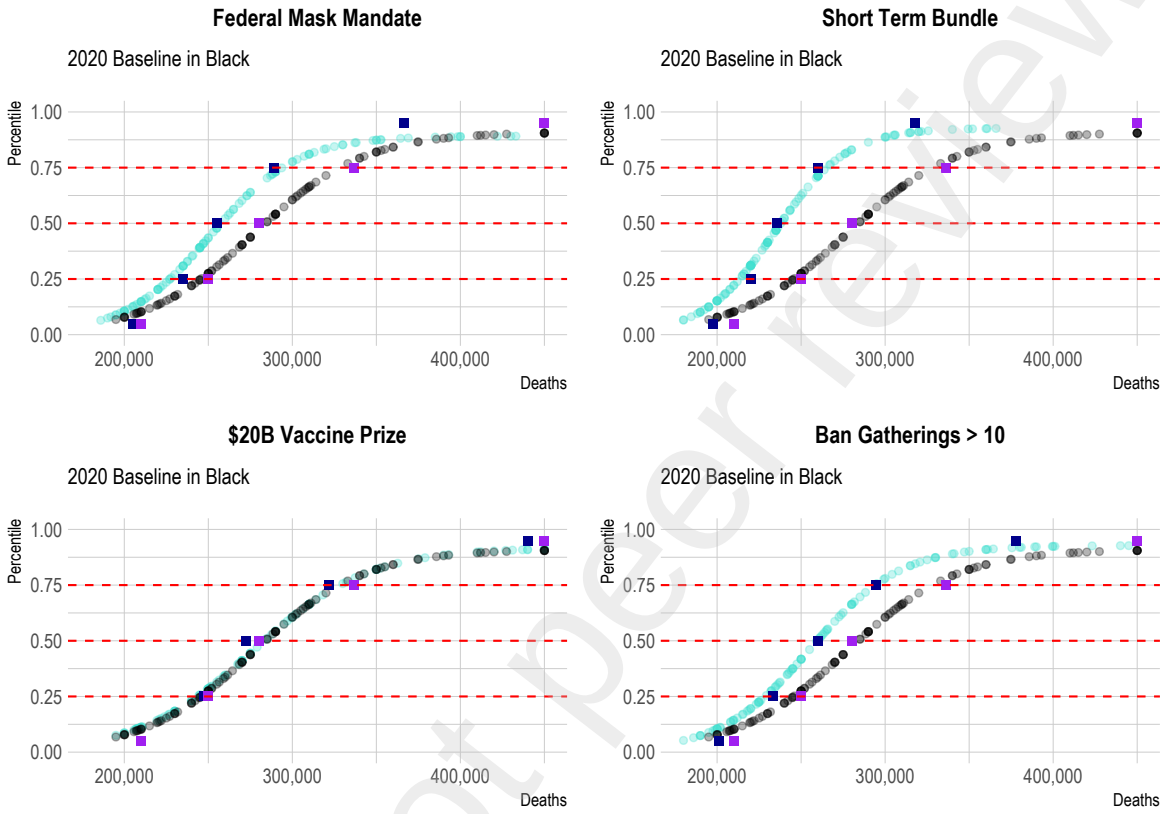
2020

Policy	**Percentile**				
	5th	**25th**	**50th**	**75th**	**95th**
Short Term Bundle	11,975	30,000	48,000	60,000	123,657
Stay-at-Home Order (3 weeks)	13,825	20,628	27,500	36,250	75,750
Ban Gatherings >10	8,625	16,250	22,250	26,500	72,000
Federal Mask Mandate	8,000	15,000	20,750	38,000	73,250
Long Term Bundle	5,381	10,438	20,000	25,000	46,500
Mask Production and Distribution	5,750	10,000	16,500	26,315	30,650
Closing Schools	5,000	10,000	15,000	20,000	32,500
$100B for Testing	3,000	8,000	13,500	16,000	25,250
Restrict Indoor Dining	6,000	10,825	10,250	22,450	30,000
Closing Universities	3,250	8,000	10,000	14,063	24,250
Federal Contract Tracing Program	5,000	5,500	8,500	8,500	14,700
Ban Gatherings >50	3,000	5,000	7,500	10,500	23,350
$20B Vaccine Prize	0	0	0	0	0

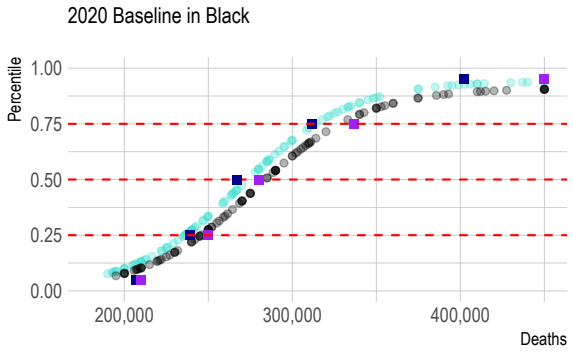
2021

Policy	**Percentile**				
	5th	**25th**	**50th**	**75th**	**95th**
Short Term Bundle	13,411	33,308	44,585	60,000	157,000
Long Term Bundle	4,961	20,808	24,712	34,000	99,500
Closing Schools	0	10,808	12,837	17,000	50,000
Stay-at-Home Order (3 weeks)	3,411	9,558	10,212	7,000	15,000
$20B Vaccine Prize	-577	10,181	5,462	10,000	35,000

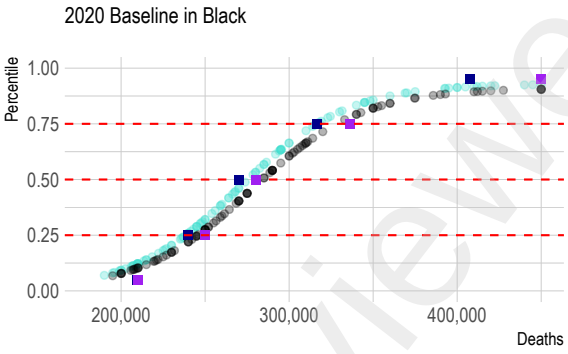
Appendix F: Imputed logistic distributions and cumulative distribution function plots of mortality reduction per policy



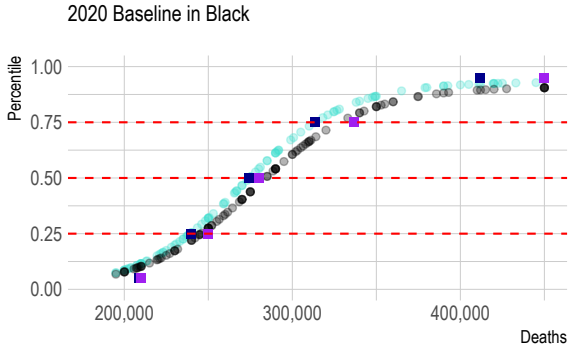
Closing Schools



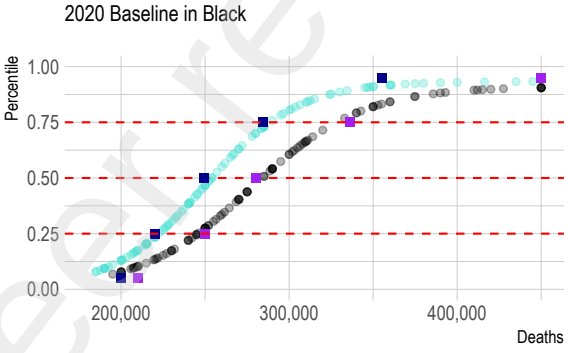
Closing Universities



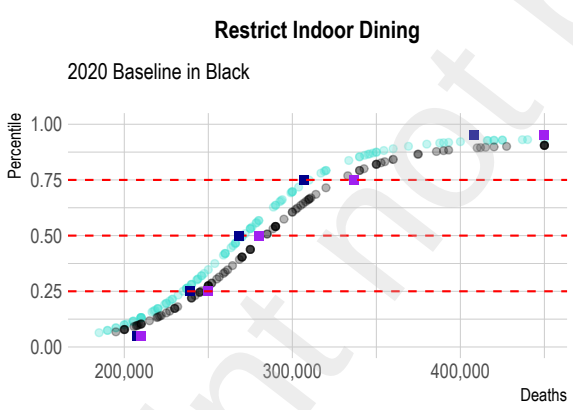
Ban Gatherings > 50



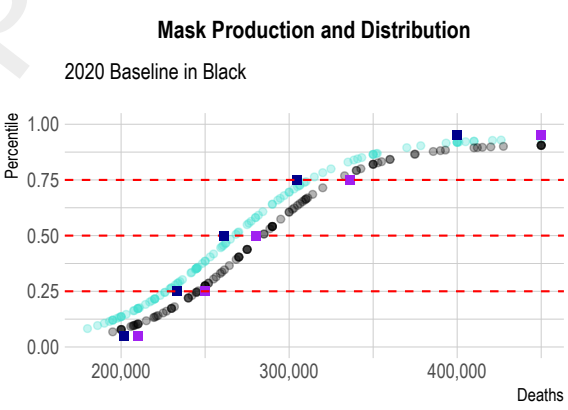
Stay at Home Order (3 weeks)



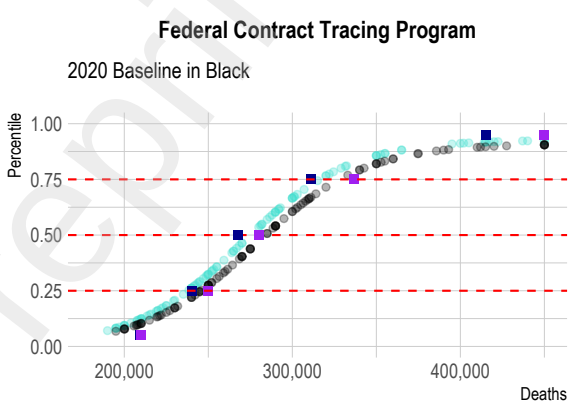
Restrict Indoor Dining



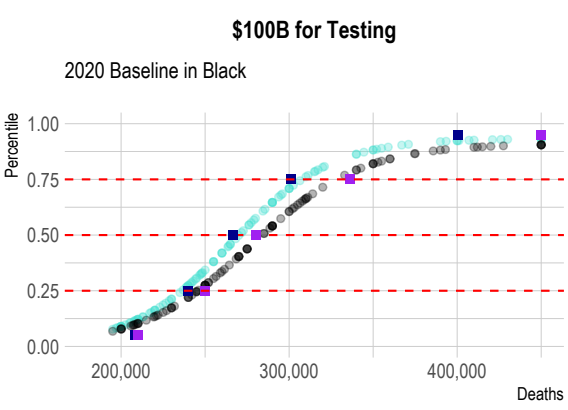
Mask Production and Distribution

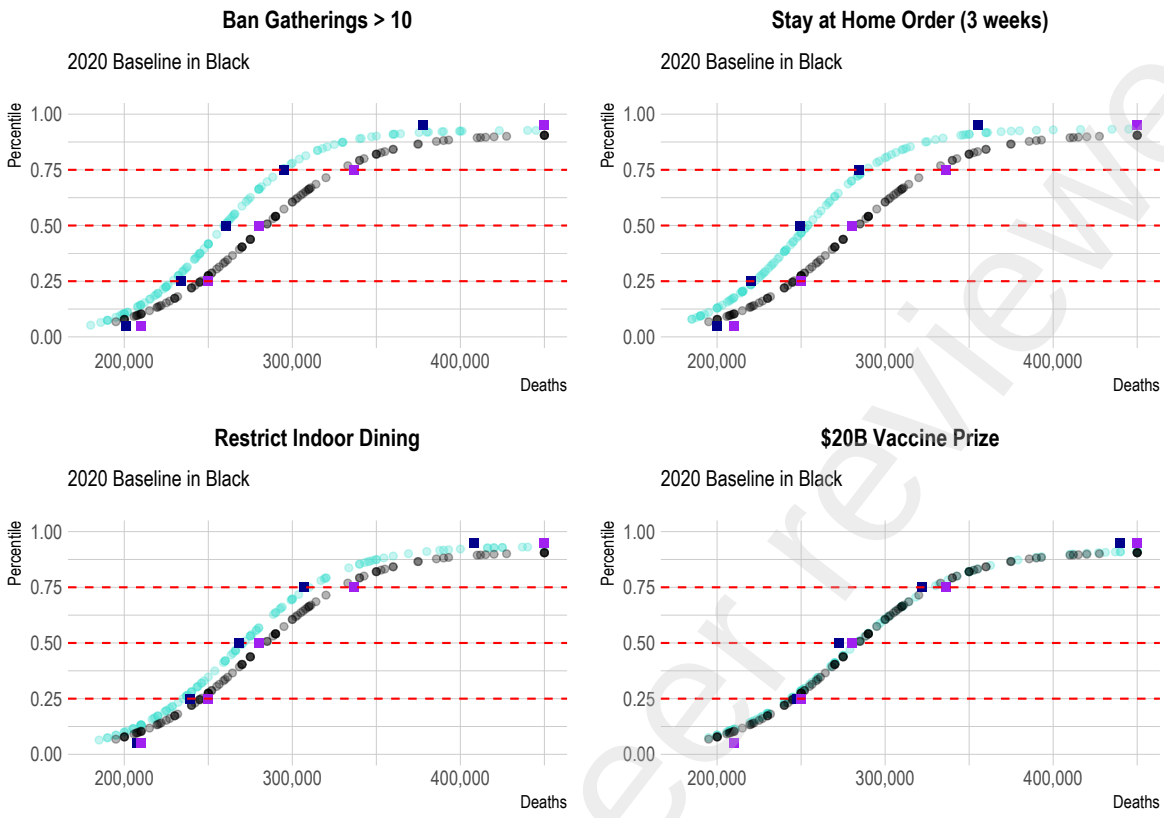


Federal Contract Tracing Program



$100B for Testing





Appendix G: Tournament post-mortem survey results

Reasons for reporting something other than true beliefs	Forecasters (of 19)
I thought that forecasters in the other group might be biased in a certain direction	4 (21%)
I thought that I might have obtained information that the forecasters in the other group did not have access to	3 (16%)
I thought that my team's median forecast was a good proxy for the other team's median, even if it was objectively inaccurate	7 (37%)
I thought I was able to identify patterns in how forecasters would respond to these questions	5 (26%)

Appendix H: Question example in Reciprocal Scoring condition

Please spend up to five minutes answering the following question:

How many total cases of COVID-19 (confirmed and probable) will be reported in the city of New York on June 30th?

The historical daily case counts for New York City can be viewed here: https://www1.nyc.gov/site/doh/covid/covid-19-data.page

Remember: the value you provide for the 5% chance must necessarily be lower than the value provided for the 25%, and so on. In other words, the sliders below must necessarily move to the right which each successive percentile.

There is a 5% chance that the number of cases will be fewer than...

There is a 25% chance that the number of cases will be fewer than...

There is a 50% chance that the number of cases will be fewer than...

There is a 75% chance that the number of cases will be fewer than...

There is a 95% chance that the number of cases will be fewer than...

Appendix I: Training process for three experimental conditions

[Screen 1: Task introduction]

Thank you for participating in this study.

We will ask some questions about events in the near future.

Before asking those questions, we will provide you with some important background information. **Please read the following pages very carefully.**

[Screen 2: Task introduction]

On the next pages, we will ask you to make forecasts about events in the near future. We will ask you to predict what the temperature will be on a certain date or how the stock market will change over the coming weeks. In total, we will ask you to forecast **10** questions.

For some of the questions, it will be very difficult to come up with a good answer without doing some research. For this reason, **you are allowed to spend up to 5 minutes researching each of the questions before answering.**

We encourage you to spend all of this time looking for information and to use your phone, computer or any other device to consult any resources you would like, including by reading newspapers, websites, or articles.

[Screen 3.1: Task introduction- *Control condition only*]

Please note that all opinions you express will be treated anonymously. The important thing is that you make judgments about the issues naturally and honestly, as you normally would if you were talking with friends about a topic and someone asked for your opinion. The researchers will not be re-contacting you to provide feedback on the accuracy of your opinions.

If you complete the 10 forecasts, you will receive $5, regardless of accuracy. Additionally, you will enter a lottery where 5 randomly chosen winners will receive a bonus of $100.

[Screen 3.2: Task introduction- *Brier Condition only*]

No one can forecast the future with complete certainty, but research shows that some people can be remarkably accurate when making forecasts. Researchers call these highly-accurate forecasters "superforecasters".

Superforecasters are people who are exceptionally good at forecasting, and who have performed in the top 2% over the course of multiple forecasting tournaments across different domains.

One of the goals of this study is to identify new superforecasters. If you are among the 5 most successful participants, you will have the opportunity to participate in future competitions as a superforecaster and **you will be awarded a prize of $100**, in addition to your compensation for completing this survey.

Note: You will be scored based on **how well your forecasts align with the actual observed outcomes for each of these questions.**

The closer your forecasts are to reality, the better your score will be, and the higher your chances will be to win one of the $100 prizes.

[Screen 3.3: Task introduction- *Reciprocal Condition only*]

No one can forecast the future with complete certainty, but research shows that some people can be remarkably accurate when making forecasts. Researchers call these highly-accurate forecasters "superforecasters".

Superforecasters are people who are exceptionally good at forecasting, and who have performed in the top 2% over the course of multiple forecasting tournaments across different domains.

One of the goals of this study is to identify new superforecasters. If you are among the 5 most successful participants, you will have the opportunity to participate in future competitions as a superforecaster and **you will be awarded a prize of $100**, in addition to your compensation for completing this survey.

Note: You will be scored based on **how close your forecasts are to those made by a panel of expert Superforecasters.**

The closer your forecasts are to those of the Superforecasters, the better your score will be, and the higher your chances will be to win one of the $100 prizes.

[Screen 4: Elicitation method]

Let's consider an example. You might be asked:

"How many **new** *cases of COVID-19 will be confirmed in the United Kingdom on the day of June 30?"*

Instead of asking for a single estimate of a number of cases on that day, we will ask you to state the numbers of cases in sentences like:

There is a 25% chance that the number will be fewer than **X** on June 30.

By typing in **X**, you are forecasting there is a 25% chance that the number of cases on June 30 will be fewer than **X.**

Another way of thinking about it is that you are forecasting a 75% chance the number will be more than **X** on June 30.

[Screen 5: Elicitation method]

For example, you will be asked to complete each of the following sentences:

There is a 5% chance the number of cases will be fewer than:_____(which means a 95% chance the number of cases will be more than that number)

There is a 25% chance the number of cases will be fewer than:_____(which means a 75% chance the number of cases will be more than that number)

There is a 50% chance the number of cases will be fewer than:_____(which means a 50% chance the number of cases will be more than that number)

There is a 75% chance the number of cases will be fewer than:_____(which means a 25% chance the number of cases will be more than that number)

There is a 95% chance the number of cases will be fewer than:_____(which means a 5% chance the number of cases will be more than that number)

[Screen 6: Elicitation method]

Important: in order to be logically consistent, **your responses to each of the questions above would have to increase from one to the next**.

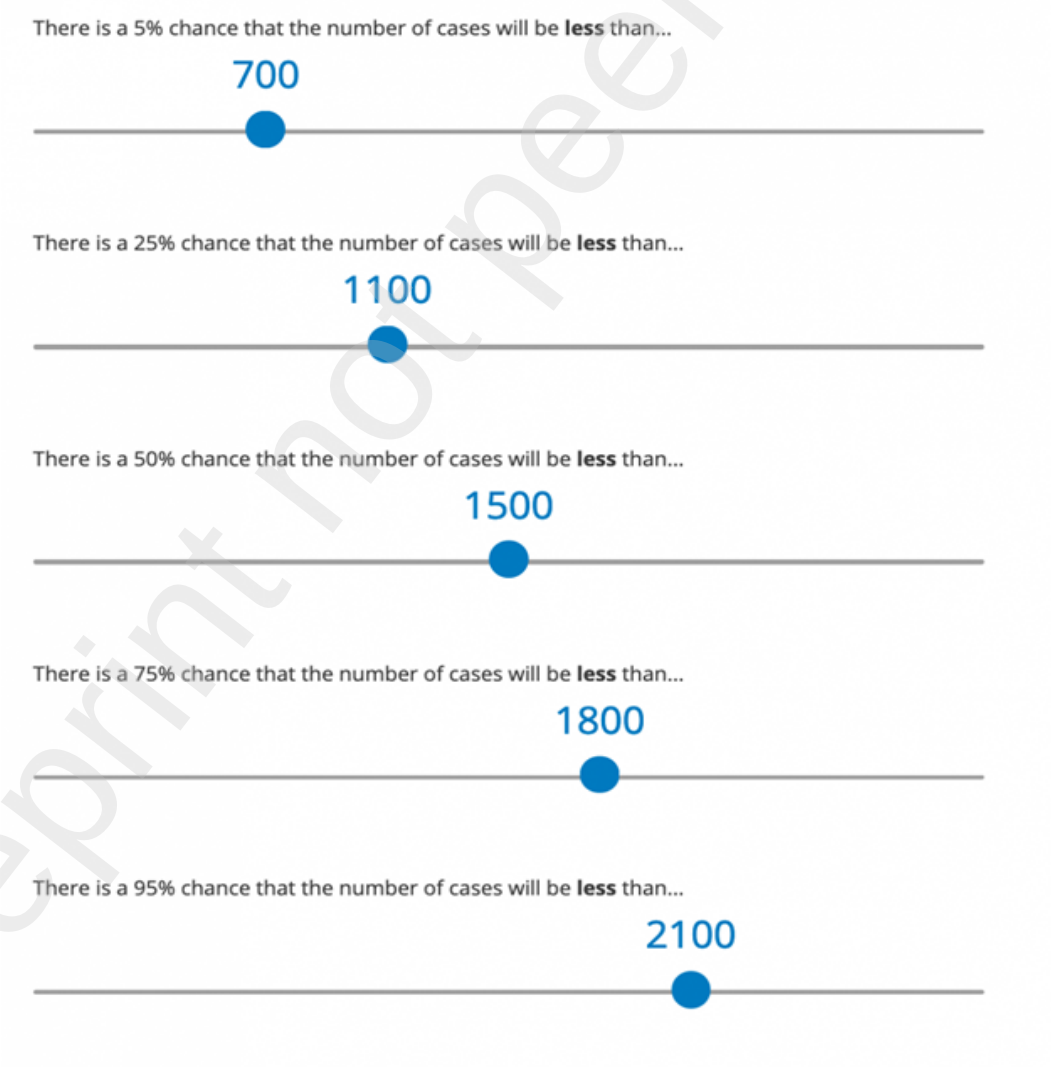
For example, if you forecast a 5% chance that new cases will be below 700,

The reason is this: Your 25% forecast must necessarily be greater than 700.The reason is this: Your 25% forecast implies a greater degree of confidence than your 5% forecast. Naturally, you have to be more confident that the correct answer will be fewer than some greater number, and less confident that it will be fewer than some smaller number. In other words, your 5% forecast (least confident) should be the lowest value and your 95% confident (most confident) should be the greatest value.

[Screen 7: Elicitation method]

We will ask you to use a sliding scale to record your answer.

For example, your answer might look like this. Notice how the blue dots representing your forecasts always move further to the right, increasing from the 5% forecast to the 25% forecast and so on:



[Screen 8: Elicitation method]

To ensure that you understand the instructions correctly, please answer the following question carefully.

Suppose you forecast a **5% chance** that the number of new UK COVID-19 on June 30 cases is going to be **fewer than 500**. You then have to forecast a number that there is a **25% chance** that the number of cases is going to be **fewer than.**Which of the following statements would be true:

A. My 25% forecast would logically have to be **greater than 500**

B. My 25% forecast would logically have to be **fewer than 500**

C. My 25% forecast would logically have to be **equal to 500**

D. My 25% forecast can **take any value**, regardless of what my 5% forecast was

[Screen 9.1: Elicitation method- ***incorrect the first time*****]**

That was not the correct answer. Please try again.

Suppose you forecast a **5% chance** that the number of new UK COVID-19 on June 30 cases is going to be **fewer than 500**. You then have to forecast a number that there is a **25% chance** that the number of cases is going to be **fewer than.**Which of the following statements would be true:

A. My 25% forecast would logically have to be **greater than 500**

B. My 25% forecast would logically have to be **fewer than 500**

C. My 25% forecast would logically have to be **equal to 500**

D. My 25% forecast can **take any value**, regardless of what my 5% forecast was

[Screen 9.2: Elicitation method- ***incorrect the second time*****]**

That was not the correct answer.

The correct answer was “My 25% forecast would logically have to be greater than 500”, since forecasts always have to increase with the level of confidence.

[A button with response requirements] I understand.

[Screen 10.1: Scoring method- *Brier Condition only*]

If you are among the 5 most successful forecasters in this study, **you will receive a bonus of $100**, in addition to your regular compensation.

To determine the prize winners, soon after June 30 we will compare your forecasts to what actually happened.

Your accuracy will be assessed used something called a "Brier Score", which gives you a more accurate score the closer your forecast is to the actual outcome. Brier scores are designed so that you do your best if you state your true beliefs about what will happen.

If you are interested, you can learn more about Brier scoring here:
https://en.wikipedia.org/wiki/Brier_score

[Screen 11.1: Scoring method- *Brier Condition only*]

If you are among the 5 most successful forecasters in this study, **you will receive a bonus of $100**, in addition to your regular compensation.

To determine the prize winners, soon after June 30 we will compare your forecasts to what actually happened.

Your accuracy will be assessed used something called a "Brier Score", which gives you a more accurate score the closer your forecast is to the actual outcome. Brier scores are designed so that you do your best if you state your true beliefs about what will happen.

If you are interested, you can learn more about Brier scoring here:
https://en.wikipedia.org/wiki/Brier_score

[Screen 12.1: Scoring method- *Brier Condition only*]

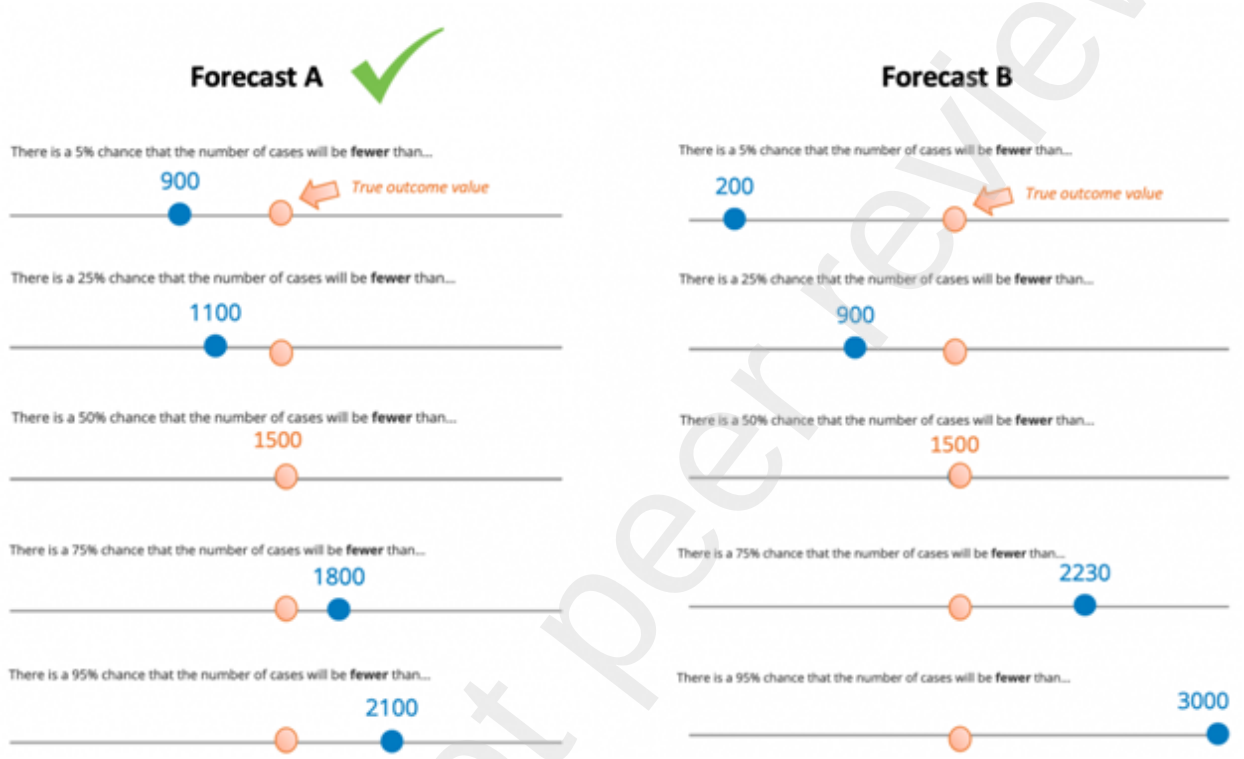
Suppose you were asked to forecast the number of **new** COVID-19 cases reported in the United Kingdom on June 30.

If the true number of cases turned out to be 1,500 on June 30, you would get a better score if you had said there was a 50% chance the value would be fewer than 1,500 than if you had said there was a 50% chance it would be fewer than 3,000.

Also, the more accurately you can forecast the upper and lower bounds of the plausible number of cases, the better your score will be.

[Screen 13.1: Scoring method- *Brier Condition only*]

Consider the two sets of forecasts below, where the blue dots represent your forecasts and the orange dots represent the actual outcome on June 30:



Even though both Forecast A and Forecast B have the same 50% forecast, **Forecast A would receive a better Brier score.** This is because Forecast A's range of plausible values was 'narrower', such that each of the forecasts were closer to the true value. In other words, Forecast A was more 'confident' that the number would not be much smaller or greater than it actually did turn out to be.

[Screen 14.1: Scoring method- *Brier Condition only*]

But while confidence is rewarded, be careful of over-confidence. Now suppose the actual number of new cases on June 30th turns out to be 800 and consider two sets of forecasts you could submit (blue dots):

Here, **Forecast B would receive the better Brier score**. This is because Forecast B assigned a 25% chance of the value being **fewer than 900** (which it turned out to be) and Forecaster A only assigned a 5% chance that this would happen. In other words, Forecast A was too confident that the number of cases would be at least 900, so they received a lower score when the number actually turned out to be 800.

Remember, because you will be evaluated using a Brier score, you will do the best if you forecast your true beliefs on each question.

[Screen 10.2: Scoring method- *Reciprocal Condition only*]

If you are among the 5 most successful forecasters in this study, you will receive a bonus of $100, in addition to your regular compensation.

A few days after you submit your forecasts, we will assess your accuracy by comparing your forecasts to the average forecasts made by a panel of 'Superforecasters.' **We will assess your accuracy before seeing the true result of the event.**

Superforecasters are people who are exceptionally good at forecasting, and who have won several large-scale forecasting tournaments across different domains. They are often among the 1-2% of the very best forecasters, and some of them do professional forecasting for a living. They are also known to be unusually good at avoiding biases in their forecasting process.

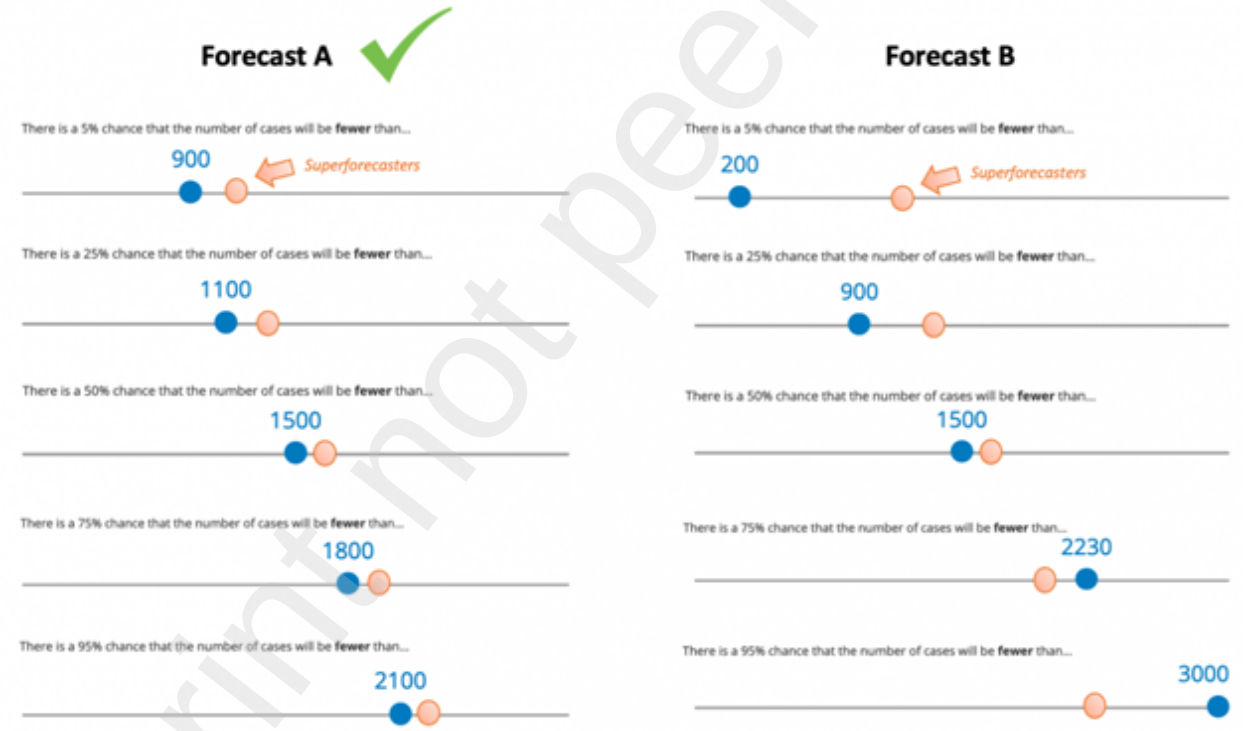
The closer your score is to that of the Superforecasters, the better your score and the higher your chance of winning a prize.

[Screen 11.2: Scoring method- *Reciprocal Condition only*]

Suppose both you and the Superforecasters were asked to forecast the number of new COVID-19 cases reported in the United Kingdom on June 30.

Once we have collected the forecasts, we will then immediately measure "how close" your forecasts were to those made by the Superforecasters, before the event occurs.

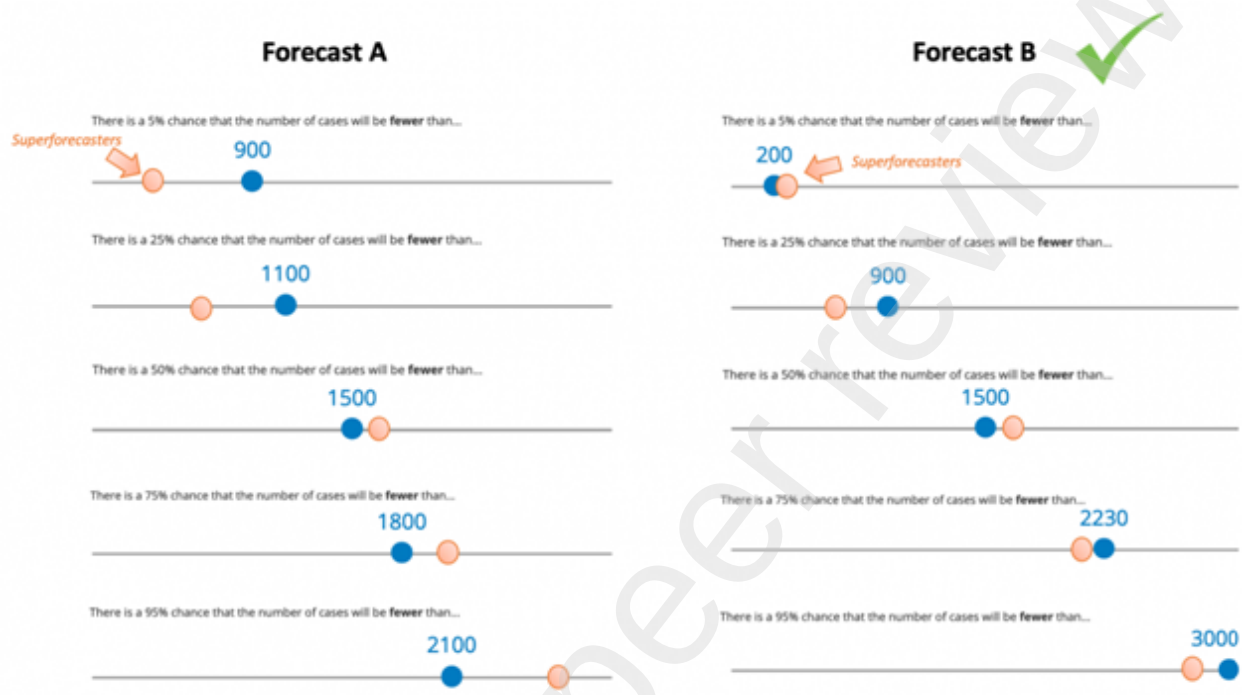
Consider the the example below, where the blue dots represent your forecasts and the orange dots represent the average forecast of Superforecasters:



Even though both Forecast A and B have the same 50% value (1,500), **Forecast A would receive a better score** than Forecast B. This is because Forecast A is closer to the Superforecaster average for the 5%, 25%, 75%, and 95% values. In other words, Forecast A was more confident that the Superforecasters would not forecast values that were extremely low or extremely high.

[Screen 12.2: Scoring method- *Reciprocal Condition only*]

But while confidence is rewarded, be careful of over-confidence. Consider the example below, where the superforecasters' forecasts are the orange dots, and your forecasts are the blue dots.



In this case, **Forecast B would receive the better score.** This is because Forecast A included a narrower range of forecasts, whereas Forecast B correctly anticipated that the Superforecasters would forecast a wider range for this question.

[Screen 13.2: Scoring method- *Reciprocal Condition only*]

Note: We will not consider how close your forecast was to the **actual** outcome, since we won't know the outcome at the time of determining the winners (on June 12th). Rather, we will **only** base your score on how close your forecasts were to those made by the Superforecasters.Because Superforecasters are known to be very accurate, the best way to maximize your chances of winning is by being as accurate as possible.Superforecasters will also be told that their score depends on how closely their forecasts match those made by other superforecasters.

[Screen 10.3: Scoring method- *Control Condition only*]

As mentioned, we would like to reward the 5 most successful participants by paying them $100 and giving them the opportunity to continue forecasting in other forecasting competitions. We may also wish to share the results online, including the names of the best forecasters.

If you are among the 8 best forecasters, do you consent to being contacted by the researchers and/or having your name shared online?

A. Yes

B. No, I prefer to remain anonymous and not to be contacted

[Screen 11.3: Scoring method- *Control Condition only and if A is selected*]

Please enter the name you would like to have shared in case you win:

(This may be your first name, full name or a username)

[A text entry]