Methods 195 (2021) 103–112

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Eliciting Beliefs about COVID-19 Prevalence and Mortality: Epidemiological Models Compared with The Street

Glenn W. Harrison [a,b,f], Andre Hofmeyr [b,c,*], Harold Kincaid [b,c], Brian Monroe [d], Don Ross [b,c,e,f], Mark Schneider [f], J. Todd Swarthout [g]

[a] *Department of Risk Management & Insurance, Robinson College of Business, Georgia State University, USA*

[b] *School of Economics, University of Cape Town (UCT), South Africa*

[c] *Research Unit in Behavioural Economics and Neuroeconomics, UCT, South Africa*

[d] *School of Philosophy and School of Economics, University College Dublin, Ireland*

[e] *School of Society, Politics and Ethics, University College Cork, Ireland*

[f] *Center for the Economic Analysis of Risk (CEAR), Robinson College of Business, Georgia State University, USA*

[g] *Department of Economics, Andrew Young School of Policy Studies, Georgia State University, USA*

ARTICLE INFO

We are grateful to the Center for the Economic Analysis of Risk for funding this research; in all other respects the funder had no involvement in the research project. We would like to thank Robert Kennedy for diligent research assistance.

Keywords:
Subjective beliefs
Beliefs
COVID-19 mortality
COVID-19 prevalence

ABSTRACT

Subjective belief elicitation about uncertain events has a long lineage in the economics and statistics literatures. Recent developments in the experimental elicitation and statistical estimation of subjective belief distributions allow inferences about whether these beliefs are biased relative to expert opinion, and the confidence with which they are held. Beliefs about COVID-19 prevalence and mortality interact with risk management efforts, so it is important to understand relationships between these beliefs and publicly disseminated statistics, particularly those based on evolving epidemiological models. The pandemic provides a unique setting over which to bracket the range of possible COVID-19 prevalence and mortality outcomes given the proliferation of estimates from epidemiological models. We rely on the epidemiological model produced by the Institute for Health Metrics and Evaluation together with the set of epidemiological models summarised by FiveThirtyEight to bound prevalence and mortality outcomes for one-month, and December 1, 2020 time horizons. We develop a new method to partition these bounds into intervals, and ask subjects to place bets on these intervals, thereby revealing their beliefs. The intervals are constructed such that if beliefs are consistent with epidemiological models, subjects are best off betting the same amount on every interval. We use an incentivised experiment to elicit beliefs about COVID-19 prevalence and mortality from 598 students at Georgia State University, using six temporally-spaced waves between May and November 2020. We find that beliefs differ markedly from epidemiological models, which has implications for public health communication about the risks posed by the virus.

1. Introduction

Beliefs that individuals hold about COVID-19 prevalence and mortality interact with efforts to manage the risks of the virus. A core concern is the relationships between these beliefs and publicly disseminated statistics, particularly statistics based on evolving epidemiological models. The COVID-19 pandemic provides an important setting to study this relationship because of the role that epidemiological models have played in public debate, and understandable biases in early editions of models that became evident over a relatively short period of time. Public awareness of the extent to which official statistics about COVID-19 in the United States (U.S.) might be biased, due to political influences and varying recording practices in different hospitals and jurisdictions, poses an additional challenge when studying this relationship. To what extent did the beliefs of individuals evolve with the forecasts of epidemiological models? To what extent did the beliefs of individuals evolve with the official reports from the Centers for Disease Control and Prevention (CDC)? To what extent did these trends affect the confidence of individual beliefs over time?

We elicit the subjective beliefs of 598 students at Georgia State University using incentivized forecasting tasks about expected COVID-19 prevalence and mortality. Our methods are designed to bracket the

* Corresponding author.
E-mail addresses: gharrison@gsu.edu (G.W. Harrison), andre.hofmeyr@uct.ac.za (A. Hofmeyr), harold.kincaid@uct.ac.za (H. Kincaid), brian.monroe@ucd.ie (B. Monroe), don.ross931@gmail.com (D. Ross), mschneider@gsu.edu (M. Schneider), swarthout@gsu.edu (J.T. Swarthout).

https://doi.org/10.1016/j.ymeth.2021.04.003
Received 30 November 2020; Received in revised form 1 April 2021; Accepted 4 April 2021
Available online 8 April 2021

range of possible beliefs that individuals have, and assess their individual confidence in those beliefs. We also developed a method that allows us to directly identify the extent to which beliefs tracked forecasts of some publicly circulating epidemiological models, quite apart from the elicitation of beliefs to address the broader questions posed above. To ensure that we were able to observe changes over time, we administered six temporally-spaced waves between May and November 2020, with different respondents selected at random for each wave.

To anchor predicted COVID-19 prevalence and mortality outcomes for the elicitation of beliefs over horizons of one month, and over horizons to December 1, 2020, we relied in part on one prominent epidemiological model, from the Institute for Health Metrics and Evaluation (IHME) at the University of Washington (http://www.healthdata.org). The IHME model has produced publicly disseminated daily forecasts of both infections and deaths throughout the course of the pandemic. We also made use of the evolving set of epidemiological models featured by FiveThirtyEight (fivethirtyeight.com), which ranged from 6 to 14 models over the course of our study, to complement the IHME model. We develop a method to partition the possible outcomes presented to subjects into intervals or *bins*, such that if a subject were to hold beliefs consistent with the epidemiological models, including allowance for statistical error, she would bet the same amount on every bin. A notable feature of our method is that it allows direct inferences about the extent to which distributions of beliefs diverge from these model-based forecasts. More extensive inferences, beyond testing this specific null hypothesis, require more structural statistical modeling, and will be undertaken in subsequent analyses, such as [1].

Epidemiological models, like any (deterministic or statistical) models, can be poor predictors of outcomes, even when designed according to accepted best practice. The IHME model that provided the basis for the baseline frame of bins we used in the study has been subject to specific criticism, with some experts arguing that its design did *not* reflect best epidemiological practice [2–4]. On the other hand, as also noted by these critics, the IHME model was the most prominently disseminated and cited source of epidemiological forecasting among the general public. This makes it a natural benchmark for our purposes. In this context it is also worth reiterating that other epidemiological models that were displayed on FiveThirtyEight informed additional frames for bins, beyond the baseline, that were used to elicit beliefs. In retrospect, the majority of these additional models did significantly outperform the IHME model in predicting U.S. COVID-19 infections and deaths.

We find that beliefs diverge markedly from the epidemiological models we used for setting bins. This finding has immediate implications for public health communication about the risks posed by the virus if we view those epidemiological models as more likely to be reliable in their predictions. With additional assumptions that allow us to infer beliefs precisely from the reports that our subjects make, we can say much more. A particularly important lesson, developed in [1], concerns the striking evolution over the 6 waves of elicitation of the level of confidence about cumulative deaths by December 1, 2020. Around August, the confidence of beliefs tightened significantly. Future research will explore the possible determinants of the changes in belief confidence, as well as any change in bias for other COVID-19 events covered by our elicitations in the U.S. and South Africa.[1]

2. Material and methods

2.1. Eliciting subjective belief distributions

The importance of eliciting subjective beliefs about uncertain events has long been clear across many disciplines. The earliest attempts to measure beliefs came from survey questions [5,6]. These have become increasingly sophisticated, with researchers now seeking to elicit whole belief distributions for non-binary events [7,8], such as the levels of COVID-19 infections and deaths that are our focus. However, surveys do not incentivize the truthful revelation of beliefs, and there is substantial evidence that using hypothetical surveys to elicit beliefs can be unreliable [9]. Our use of an incentive-compatible mechanism to elicit beliefs makes our approach fundamentally different to survey responses, and more informative.[2] The concept of subjective belief was formally developed in economics and decision theory as an extension of the notion of revealed preference [11]. Just as the strength of preferences for fine wine over plonk can be revealed by *purchase* decisions when the relative prices of the two types of wine are varied, beliefs can be revealed by *betting* decisions that depend on a particular outcome, such as the level of COVID-19 infections, *reported* by a certain source, such as the CDC, on a specific day in the future.

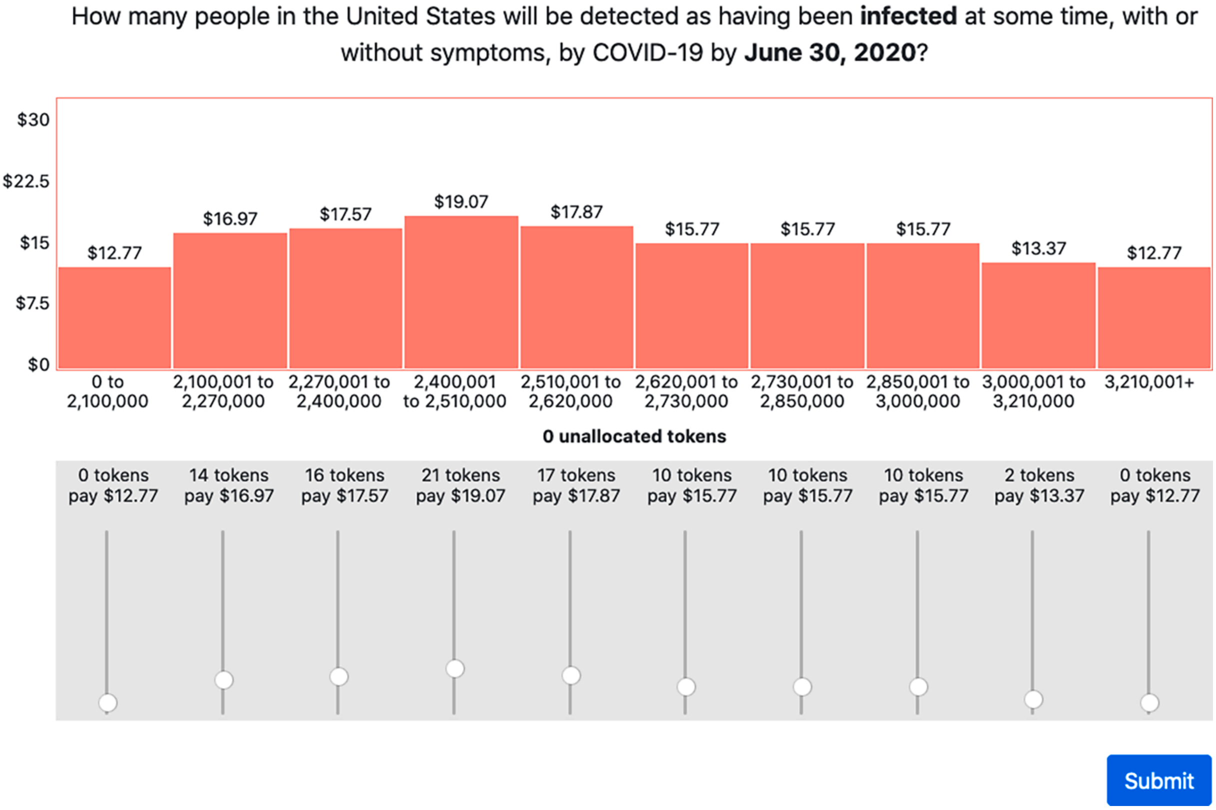
A key development in the reliable elicitation of subjective beliefs was operationalizing this notion of beliefs revealed by betting, by observing changes in betting decisions as the relative odds offered by bookies are varied. Imagine an array of bookies, lined up in terms of their odds that the COVID-19 infection rate will go up in the next month, rather than stay the same or go down. Some bookies offer great odds that it will go up, and some offer great odds that it will go down, and there are many bookies in between. Now allow someone to place a bet of $1 with each bookie. If the bettor is risk neutral, the point at which they switch from betting that infections will go up to betting that they will not go up tells us the odds that this person places on these events, and from those odds we can infer the person's subjective probability of infections going up.

It is a small formal step to present this array of bookies in the form of a "scoring rule," which translates different bets into payoffs for the bettor, depending on the realized outcome or event [12,13]. And in turn we can generalize these ideas to placing bets on several events, such as the event that infection rates go up by more than 1 percentage point, the event that they go up by between 0 and 1 percentage points, and the event that they go down. In this way we can elicit the subjective probability mass function over these events, or indeed the probability distribution function for continuous events [14]. Or we can divide a continuous event, such as the level of COVID-19 infections by June 30, 2020, into 10 bins that partition the event space over which we seek to elicit beliefs, as in the experiment we report. And since we are asking people to place bets with simulated bookies, with varying odds defined by a scoring rule, this is easy to do with real money, and thereby provide incentives for truthful revelation of beliefs *cum* bets by using "proper" scoring rules [15].

Our method is intended to be general. Consider a policy setting in which a statistical model provides predictions about macroeconomic outcomes, and policy-makers base their recommendations on those predictions. Most statistical models, particularly in economics, rely on some data that are collected with a lag, and with data that often undergo major revisions over time. Invariably, senior decision-makers come to make decisions armed with predictions from a model that they know misses some information. It could be that predictions made today from the statistical model are conditioned on interest rates or exchange rates that applied a month ago, since all other data needed for the model has a one-month lag in collection. But the decision-makers know that current interest rates or exchange rates have changed sharply in the last few weeks. In this setting the subjective beliefs of the decision-makers are formed by some combination of the statistical model and their beliefs about how that knowledge about the recent past affects the predicted macroeconomic outcomes. And even if the predicted macroeconomic outcomes are expected to be the same, knowledge of actual outcomes in recent weeks might affect the confidence intervals around the

[1] We conducted a parallel, multi-wave experiment using the same elicitation methods in South Africa, although that is not a focus of our attention here.

[2] See [10] for an example of the use of survey questions to elicit beliefs about COVID-19.



How many people in the United States will be detected as having been **infected** at some time, with or without symptoms, by COVID-19 by **June 30, 2020**?

0 unallocated tokens

0 tokens pay $12.77
14 tokens pay $16.97
16 tokens pay $17.57
21 tokens pay $19.07
17 tokens pay $17.87
10 tokens pay $15.77
10 tokens pay $15.77
10 tokens pay $15.77
2 tokens pay $13.37
0 tokens pay $12.77

Fig. 1. Subjective Belief Task Interface and Bets of Subject #183 on May 29, 2020.

predictions.[3] A comparable challenge regularly faces Chief Risk Officers and their forecasts of major financial risks.[4]

Fig. 1 is a screenshot of the experimental software we developed to elicit the beliefs of each subject about COVID-19 prevalence and mortality. This subjective belief question was presented to subjects during Wave 1 of our study, which took place on May 29, 2020. Fig. 1 shows the actual bets, in the form of a token allocation, of subject #183, and the amount to be paid depending on the answer to the question. The answer was verified using the first public report provided by the CDC *after* the date in the question, which was explained to subjects through audio-visual instructions before they completed the task.

Armed with probability mass functions over ten events, as represented in Fig. 1, which characterize subjective belief distributions over the levels of COVID-19 prevalence and mortality, we can analyse the bias and confidence of those beliefs. Bias is just the familiar concept from statistical estimation: how different is the weighted average belief from the realized event, or the best available econometric or epidemiological model at the time [17], or the claims of prominent media or political leaders? All of these types of “target beliefs” to assess bias are actually useful metrics for different reasons, so there is not just one measure of bias that is of interest.[5] Confidence is just the familiar concept of imprecision from statistical estimation, most commonly captured by the variance of beliefs about their mean. We prefer to think of confidence more broadly to reflect the variability of beliefs, so we can also consider skewness and kurtosis, but the point is to pay attention to more than just the weighted average or mode of beliefs. One can only characterize bias and confidence if one elicits subjective belief *distributions* [18], which of course allow for the special case of degenerate beliefs held with

[3] In stylized form, this is exactly what happens in the opening hours of the important Federal Open Market Committee meetings of the U.S. Federal Reserve every month. Forecasts of the future economy have been distributed by staff of the Federal Reserve Board of Governors, and the discussion leads to a consensus as to what the Committee believes is likely to happen to the economy. Based on that consensus, critical policy decisions by the voting members of the Committee are made [16]. The consensus might be the same as the forecasts of the statistical model, it might be different in expectation, or it might just be different in terms of confidence intervals. Our method may be framed as a formal way to characterize how these initial statistical forecasts compare to the views of the Committee members. In effect, it would have the Committee members place bets, with proceeds to a worthy charity of course, on the future outcomes of certain key macroeconomic variables.

[4] A comparison of models and beliefs, similar to ours but for Chief Risk Officers (CRO) and a statistical model, was reported by [17]. In that case the predictions of the statistical model for a one-year horizon, generated just prior to the belief elicitation, were used to calibrate a belief elicitation task presented individually to the CRO subjects. Comparisons of their predictions suggested, *inter alia*, that the CRO predictions did not have the extreme “tails” of the statistical models.

[5] Our subjects were literally rewarded for their beliefs about the report by the CDC of the cumulative level of infections or deaths from COVID-19 on a certain date. This report, of course, was an estimate. Such reports were often revised over time, as more data, better data, and alternative methods of estimation were employed. Hence we refer generally to a “target” value against which elicited beliefs were compared, rather than some true, objective value.

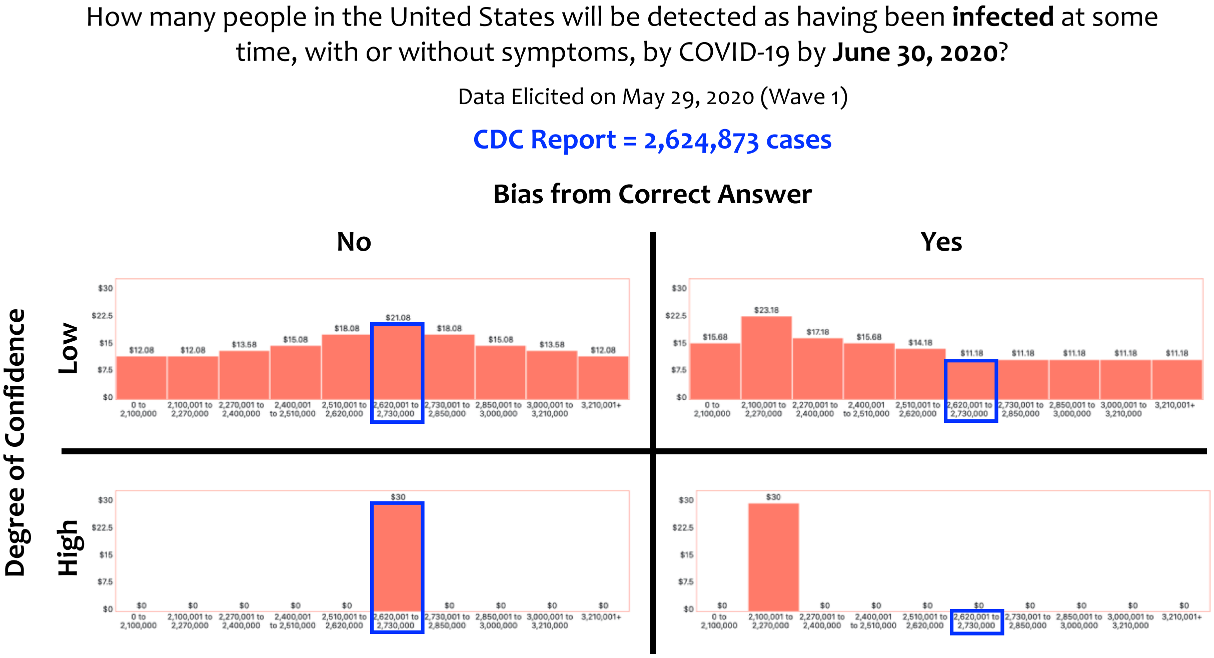


Fig. 2. Bias and Confidence of Subjective Belief Distributions

certainty.[6] Fully Bayesian epidemiological models of COVID-19 infections and deaths provide posterior predictive distributions of future levels, which can be used to also make determinations of whether subjective beliefs are “overconfident” or “insufficiently confident,” using the approach documented in [20].

Our approach here to our subjects’ elicited beliefs is Bayesian. We compared elicited beliefs with expert epidemiological opinion. The latter might be taken as reasonable priors which agents use to produce their own posteriors given what individual data they have. Also, we describe beliefs as a probability distribution over outcomes, as a Bayesian would. Priors here refer to already held beliefs about the probability of some statement; in the context of medical testing, priors about having a disease might start from the population base rate. Posteriors are revisions of those beliefs formed after obtaining new evidence, such as a positive disease test. However, we make no explicit use here of Bayesian updating from priors plus data to explain changes in posterior beliefs of our subjects from wave to wave.

Fig. 2 shows the realized answer, as reported by the CDC, to the question from Fig. 1, and hypothetical bets that vary according to whether they are biased relative to the number of infections by June 30, 2020, and the confidence with which these beliefs are held. Per the experimental protocol, the official reports from the CDC are treated as the correct answer that determine subject payments. The top left quadrant of the figure represents an unbiased, but relatively low confidence, set of bets, in the sense that the largest bet was placed on the correct answer, but bets were also made on other events. The bottom left quadrant also represents unbiased beliefs, but held with a degenerate level of confidence in the sense that all tokens were bet on the correct event. The two right quadrants represent biased beliefs because no tokens were allocated to the correct event, but clearly differ according to the strength with which beliefs were held.[7]

A direct implication of incentivizing bets with a proper scoring rule is that if someone believes that each event, as represented by the bins in a task, is equally likely to occur, the person will bet exactly the same amount on each bin, as represented in Fig. 3. Thus, when someone bets anything other than the same amount on every bin, this reveals that they do not consider every event as equiprobable. We constructed bins over which to elicit beliefs about the number of infections and deaths due to COVID-19 in the U.S. either one month in the future or by December 1, 2020. These bins were constructed such that if a person’s bets differ across bins, this non-uniformity across bins reveals that the person’s beliefs deviate from epidemiological models of infections and deaths due to COVID-19.

The first step in constructing these bins is to define the distribution of underlying events. We assumed that deaths and infections, scaled to the population of the U.S., follow a Beta distribution. The Beta distribution is flexible enough for our purposes, has well-defined higher moments, and finite support over an interval. This last property ensures that the number of people who will be infected or die due to COVID-19 cannot be negative or greater than the U.S. population. In addition, the Beta

[6] The notion of bias is used in several different ways across various disciplines. In statistics and econometrics it typically refers to an estimate of some parameter, such as the average of an estimated belief distribution. In this case, one would construct tests that compare the point estimate of the parameter to the “target” estimate, using the estimated standard error of the point estimate. An alternative approach that is standard for Bayesians is to define some “region of practical equivalence,” or ROPE, that describes differences between the parameter estimate and the “target” estimate, and then compare that ROPE to the highest density interval (HDI) of some estimated distribution. In our case the ROPE is the distance between the average belief and the target estimate, and the HDI is defined over the elicited subjective distribution of beliefs (*not* the distribution of the mean as a parameter estimate). For symmetric distributions the HDI is the familiar equal-tailed interval. To an economist, the ROPE refers to the bias that is of economic significance. To a Bayesian, the ROPE allows a natural statement of what classical statisticians mean by the testing of a point-null hypothesis [19], by turning it into an interval hypothesis appropriate for the inferences at hand. See Appendix A for details on the ROPE we constructed for our Bayesian statistical analyses.

[7] We used a quadratic scoring rule (QSR) to incentivize truthful revelation of beliefs. As a proper scoring rule, the QSR provides the highest *expected* reward if risk neutral subjects report their true beliefs, and therefore penalizes subjects for betting on events to which they do not assign positive probability. Unless a subject reports degenerate beliefs, as in the bottom left or right quadrant of Fig. 2, the QSR still provides payment for bins to which no tokens have been allocated, as in the top left or right quadrant of Fig. 2.

distribution is well suited to characterizing the bias and confidence of subjective belief distributions. Finally, it has two sufficient statistics, and therefore the shape of the distribution can be defined by two points, or *anchors*, along its cumulative distribution function (CDF), if the cumulative density at each anchor is known or imposed.

We therefore set out to define pairs of anchors that consist of a lower anchor, such that there is a probability of 0.1 that the true statistic would be less than this amount, and an upper anchor, such that there is a probability of 0.2 that the true statistic would be greater than this amount.[8] For each pair of anchors, a Beta distribution was defined that uniquely satisfies these two sufficient statistics. This Beta distribution was then used to define 10 bins such that each bin represented 10% of the full distribution's cumulative density.[9] This bin construction exercise ensures that if a person's beliefs were the same as the distribution we defined, which was based on epidemiological models, they would maximize their expected earnings and their expected utility by betting the same amount on each bin.

Our method for designing which bins to present to subjects was intended to provide general information about the beliefs of individuals that reflected our hyper-priors about the underlying data generating process. Our method also served to generate a sharp, direct test of the specific null hypothesis that the beliefs of individuals tracked those informed by epidemiological models. And by the beliefs from individuals and the models "tracking" each other, we mean much more than aligned weighted averages: we insist that they also track each other in terms of levels of confidence. This additional criterion allows us to determine if the evolution of epidemiological understanding and modeling, which was dramatic during the period of our elicitation, is matched by an evolution of individual beliefs.

2.2. *Reflecting epidemiological models*

To effect this test of the null hypothesis, we need some characterization of the beliefs that might be arrived at by attention to "epidemiological models." To do that we started with the IHME model, and used the forecasts that it provided to generate the bins we refer to as frame 0.[10] And, more specifically, our method generated bins that implied equal weight should be given to each bin, in terms of bets implemented with token allocations. Proper scoring rules incentivize the truthful revelation of beliefs of risk neutral bettors. There are deep theoretical, experimental, and statistical issues that arise when agents are *not* risk neutral, because then they can make bets to hedge against risk [15]. For example, an *extremely* risk averse decision-maker might bet the same amount on every bin in a subjective beliefs task to ensure zero variance in payment, regardless of the event that is realized. Thus, if the subjects in our experiment all bet the same amount on every bin, we would be unable to directly infer whether this was due to high levels of risk aversion or to beliefs that are consistent with epidemiologically informed forecasts.[11] However, to the extent that subjects do not bet the same amount on every bin, this implies that their beliefs are not consistent with the epidemiologically informed forecasts *regardless* of their levels of risk aversion. This property is a powerful innovation in the method developed and applied here: it is apparent that risk preferences of individuals *only* matter if subjects do not bet the same amount on every bin. Hence we are able to test this null hypothesis by directly comparing the token allocations we observe from individuals, without any need for adjustment for their risk preferences.

The specific epidemiologically informed model used for frame 0 was then used as the basis for adjustments to generate the bins reflected in frames 1, 2, and 3, which were also informed by consideration of additional publicly circulating epidemiological models. Apart from allowing us to test for wholesale deviations from the hyper-priors reflected in frame 0, these frames themselves can be viewed as reflecting beliefs informed by wider ranges of epidemiological models. Our method then adds the constraint that someone holding beliefs consistent with those models would bet exactly the same amount on every bin. Thus, frame 0 gives greatest prior weight to one specific epidemiological model, and frames 1, 2, and 3 use bins reflecting alternative ranges of wider epidemiological modeling. In this sense, our complete set of frames is designed to reflect "epidemiological models" as a whole, respecting the inevitable changes in the number of such models available for public scrutiny, and modeling assumptions of different experts, over the course of the pandemic.

Our method to select two anchors for our belief elicitation required us to focus on prospective outcomes for COVID-19 statistics at various future time points, over which subjects could then place bets. We aimed to base these anchors on credible epidemiological models, presenting us with several challenges.

First, when COVID-19 was declared a pandemic, epidemiologists still had relatively little knowledge of its transmission vector, but this knowledge improved rapidly and steadily over the course of our study frame [24]. This improvement of knowledge resulted in changes in the specification structure of models as our study unfolded, and the addition of new models that were made available between waves of our study.

Second, no single model supported forecasts of all of the outcomes on which we asked subjects to report beliefs. So we were forced to sacrifice some consistency with respect to the set of epidemiological models considered over time, as well as across outcomes at a point in time.

Third, on some of our waves the virus was spreading very quickly and there were lags between CDC reports and model forecast updates. On some occasions, when the CDC incorporated retrospective data from heavily affected states as jump shifts, the effects of these lags were substantial. Statistical efficiency implied that we not present subjects with bets on outcomes at the lower end of infections or deaths that had already become impossible. In general, our method was to use our own hyper-priors to construct a specific null hypothesis. This entailed using as much information about the pandemic as was available to us, rather than devising a procedure for mechanically applying epidemiological models. At the same time, anchoring our hyper-priors on epidemiological forecasts in a consistent way was also a crucial element of our method.

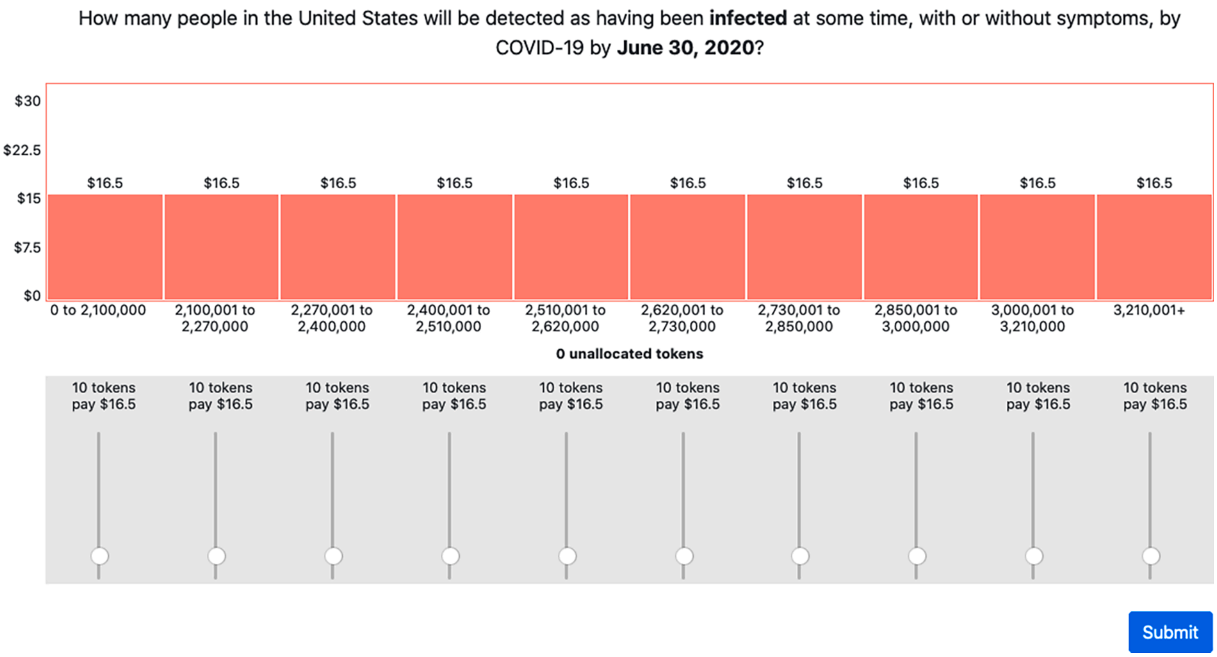
In the face of these and other challenges, we adopted the following approach to selecting belief elicitation anchors for COVID-19 prevalence

[8] Section 2.2 discusses the bin anchoring calculations we performed for each wave of the study.

[9] In general, any parametric distribution with a defined CDF and **S** sufficient statistics can be defined by **S** points along the CDF. Let F(x | α, β) be the cumulative density of the Beta distribution below x, with shaping parameters α and β. For each frame, we pick (x, y), such that F(x | α, β) = 0.1, and 1 - F(y | α, β) = 0.2. We then combine these two equations such that h(α, β) = F(x | α, β) + F(y | α, β) – 0.9 = 0, and solve for the unique roots α*, β* such that h(α*, β*) = 0. Finally, we use the resulting Beta distribution F(x | α*, β*) to define the bins for that particular frame.

[10] See [21] for a review of the historical context, modeling assumptions, accuracy, and criticisms of the IHME model. See [22] for an early discussion of susceptible-infected-recovered (SIR) model validity given human behavior in response to the pandemic. Finally, see [23] for an evaluation of COVID-19 models over time.

[11] During the experimental session, we also elicited the risk attitudes of each subject to account for the possibility of hedging in the elicitation and estimation of subjective beliefs. We do not focus on risk attitudes here, because they are unnecessary for our inferential objective.



How many people in the United States will be detected as having been **infected** at some time, with or without symptoms, by COVID-19 by **June 30, 2020**?

Fig. 3. Bets for Equiprobable Events.

and mortality in the U.S. population.[12]

To begin, for each of our six waves we selected a *baseline distribution* (BD0) based primarily on forecasts from the IHME. We selected this model because among those that were available from the beginning of our study, it uniquely provided specific projections of both case and death numbers for every future date through our planned time course. However, we allowed adjustments to this distribution wherever it was incompatible with the actual incidence of infection and mortality, due to lags. During moments of rapid transmission, mortality reports were a basis for estimates of infections that were more reliable than direct infection reports themselves, and our method called for this information to be incorporated into our hyper-priors.

We did not limit anchors to the adjusted IHME-driven baseline distribution for three reasons. First, we sought to reduce the likelihood that many subjects might place all of their tokens in one extreme bin or the other, thus failing to provide us with much information about the distribution of their beliefs. Second, we aimed to avoid being limited to broad bin ranges that would fail to provide subjects with opportunities to report relatively precise beliefs if they indeed held such beliefs. Finally, we did not want to end up with uninformative responses, which could occur with very wide bins if subjects bet all of their tokens on the same bin.

We therefore constructed three additional bin anchors (BA1 - BA3) that shifted forecasting anchors relative to the baseline. To connect these to expert observation and modeling, we drew information from additional epidemiological models. The data journalism website FiveThirtyEight consolidates models produced by leading public health research institutes. The number of these reported models varied during the course of our study, from 6 on Wave 1 to 14 by Wave 6. We used the mortality forecasts of these additional models to constrain construction of bin anchors BA1 - BA3 for each wave. Since these models, unlike the IHME model, do not forecast infections, when anchoring bounds for infections we imposed the case fatality rate (CFR) that prevailed at the time of the wave according to the CDC. We then assumed that this rate would converge linearly over time to the CFR of the IHME model for December 1, 2020; again, the IHME model provided the only long-range forecast available during early waves.

We established anchors for BA1 in each wave by replacing the mortality anchor for BD0 by the bottom of the forecast range for the most “optimistic” model in the FiveThirtyEight suite as of the wave in question, where “optimistic” means the model that forecast the lowest number of deaths consistent with actual mortality reports on the day before launch of the wave in question. The upper anchor was then adjusted so that the probability density function (PDF) would replicate the baseline distribution BD0 as closely as possible, subject to the constraint imposed by the assumption made above about the CFR value for each wave. Setting bottom anchors for each BA1 was the step in the construction most likely to require *ad hoc* adjustment due to lag effects. In such instances, basing the top anchor on the most “optimistic” model in the suite required us to relax the assumption of a uniform PDF. To avoid suggesting implausibly over-precise estimates to subjects, such as 50,123 deaths, all anchors were converted to integers rounded to the nearest multiple of 10.

We constructed anchors for BA2 by replacing the upper anchor of BD0 by the upper end of the most “pessimistic” model in the FiveThirtyEight suite, and adjusting the bottom anchor by analogous restrictions as for the BA1 construction above. Where lag effects required adjustments to bottom anchors on BD0, corresponding adjustments were made to bottom anchors of BA1 by reference to the assumed CFR for that wave.

Finally, we constructed anchors for BA3 by setting the upper anchor to the top of the error range of the implied BA2 model for $p = 0.05$, then shifting up the bottom anchor by again maintaining the PDF of BD0 constrained by the assumed CFR for that wave.

Thus, the ranges presented to study subjects were based on one set of bin anchors (BD0) that treated the IHME forecast as if it were the most informative, one set of anchors (BA1) shifted in an “optimistic” direction that remained within the range of expert forecasts and actual reports as of the day preceding the wave, and two sets of bin anchors (BA2 and BA3) shifted in a “pessimistic” direction, but also within the bounds of epidemiological modeling. The motivation for this asymmetry between

[12] We also asked subjects to forecast prevalence and mortality rates among Americans aged 65 years and older, in light of the crucial role of their far higher mortality in driving policy responses. The construction of anchors for this part of our experiment involved special problems due to progressive decline in available data quality over the course of our study. This aspect of the overall project will be discussed elsewhere.

optimistic and pessimistic representations reflected the fact that lag effects sometimes violated optimistic, but never pessimistic, bounds of distributions.

With this set of bin anchors for prevalence and mortality statistics over one-month and December 1, 2020 timeframes we used our Beta distribution algorithm to partition the event spaces and define the set of bins for the task. Four sets of bin anchors produced four sets of bins per belief question, which we refer to as the *frames* for that question: BD0 defined the anchors for frame 0, and BA1, BA2, and BA3 defined the anchors for frames 1, 2, and 3, respectively.

One frame per belief question was drawn randomly for each subject, so frames varied between subjects in the task. The construction of these frames allows us to draw inferences about the extent to which non-expert subjective beliefs differ from expert forecasts encoded in epidemiological models to the extent that bets vary from bin to bin.

3. Results

We focus here on subject bets implemented by token allocations for the one-month forecasting horizon across waves 1–6 of our study. We limit our analyses to this horizon for ease of exposition. This sample consists of 598 subjects across the six waves. Figs. 4 and 5 represent the data from the 112 subjects who took part in Wave 1.

Fig. 4 shows the distribution of token allocations from May 29, 2020 (Wave 1) for the number of COVID-19 infections in the U.S. by June 30, 2020. The distributions differ markedly across frames, which suggests that the way in which event spaces are anchored and partitioned affects subjects' token allocations, even though each set of anchors was consistent with epidemiological models of the pandemic. However, to draw valid inferences about differences across frames both with respect to each other and the number of cases reported by the CDC, it is essential to account for the risk attitudes of subjects [15]. This is not the focus of our analyses here. Despite these apparent differences across frames in Fig. 4, the crucial result is that subjects did not bet the same amount on every bin. To test this hypothesis we estimated a Bayesian model of an ordered logit data generating process with appropriately diffuse priors; see Appendix A for further details. We defined the null hypothesis in terms of inferred posterior probabilities for each bin between 0.9 and 0.11 for our sample size. This "region of practical equivalence" (ROPE) from a Bayesian perspective [19] corresponds to the range of posterior estimates generated from randomly selected token allocations for each bin between 0 and 20. The posterior probability of the data in Fig. 4 being in this interval is less than 0.001, calculated over all frames and waves.[13] Thus, subjects did not bet the same amount on every bin, which is a necessary condition for the beliefs of subjects to be consistent with epidemiological models of the spread of the virus. If some of the token allocation distributions were more flat than others, this would suggest that the epidemiological modeling associated with that frame was more closely aligned with the beliefs of subjects, but clearly no such inference is valid on the basis of the distributions in Fig. 4.

Appendix B shows the distribution of token allocations elicited in waves 2–6 of our study of the number of COVID-19 infections in the U.S. one month after the date of each wave. While there are some interesting differences across waves, which reflect the (rapid) evolution of the pandemic in the U.S., better scientific understanding of the spread of the virus, and the proliferation of epidemiological models that had more data to feed their predictions, the overall pattern is the same: subjects' beliefs differ significantly from epidemiologically informed models of COVID-19 infections.

Fig. 5 shows the distribution of token allocations from May 29, 2020 (Wave 1) for the number of COVID-19 deaths in the U.S. by June 30, 2020. Unlike infections, the distributions across frames are similar, but formal tests of the extent to which they differ require adjustments for risk attitudes. Again, the crucial result for our purposes is that subjects did not bet the same amount on every bin. The Bayesian posterior probability of this null hypothesis is, again, less than 0.001 over all frames and waves. Thus, the beliefs of subjects about COVID-19 deaths are not consistent with epidemiological modeling.

Appendix C shows the distribution of token allocations in waves 2–6 of the number of COVID-19 deaths in the U.S. one month after the date of each wave. Differences across waves are less pronounced in comparison to beliefs about COVID-19 prevalence, but subjects' beliefs clearly differ from epidemiologically informed models of deaths attributed to the virus.

4. Discussion

A general challenge implicit in our design was that the U.S. has not, as we write, yet implemented large-scale randomized testing for COVID-19.[14] Consequently, detected cases involve over-representation of infected people who presented with morbid symptoms. Furthermore, accurate tracking of prevalence and mortality in the U.S. has been impeded by decentralized administration and politicized conflict [25]. Epidemiologists universally acknowledge that undetected cases with lower morbidity outnumber detected cases [26]. The implication is that the evidence-based forecasting in which we asked our subjects to engage was not directly of the disease itself, but rather of the evolution of the processes used by public health officials to arrive at announced statistics and projections. It is open to question to what extent people behaviorally manage their health risks by responding to expert forecasts, and to what extent they choose behavior on the basis of their own idiosyncratic representations of diseases.

Coupled with these issues were significant changes in scientific understanding of the virus over the time period of our study, which presumably also influenced the risk mitigation efforts of individuals. These changes in expert understanding can be summarized as follows: estimations of the frequency of fomite transmission declined; estimations of the frequency of aerosol transmission increased; estimations of the efficacy of widespread mask use against prevalence, morbidity, and mortality increased; estimations of the weight of behavioral responses, independent of public-health policy choices, increased; and estimation of the extent of path-dependence in transmission geography due to "super-spreading" events increased. While our study does not speak directly to this improving scientific knowledge of the virus, the fact that we constructed a pseudo panel of participants means that we can track the evolution of beliefs about COVID-19 prevalence and mortality over time [1]. This will allow us to determine whether beliefs became more or less biased, and whether the confidence with which these beliefs were held varied, as more information about the virus became available. We will proceed with this line of investigation in subsequent analyses.

Figs. 4 and 5, together with the complementary figures in Appendices B and C, show that forecasting COVID-19 infections is fraught with difficulty, certainly in comparison to deaths. We define the "correct" answer for our subjects as meaning "correctly matching the CDC's estimated report." Fig. 4 shows that this correct answer in frame 0 about the level of infections on June 30, 2020 fell into the last bin of the event

[13] The same conclusion applies if we only examine frame 0 (reflecting the IHME model), only examine frames 1, 2 and 3 (reflecting all models other than the IHME model), or only examine frames 2 and 3 (reflecting the best performing models from an *ex post* perspective). The same conclusions apply for beliefs about deaths as well as beliefs about infections, except for deaths in frame 0, wave 6, where the Bayesian posterior probability is less than 0.01.

[14] We refer here to randomized testing using the entire population as the sampling frame. Due to the clear existence of many asymptomatic cases, randomized testing of only people who present with symptoms, as has very helpfully characterized the public health response in a number of countries such as Germany and South Korea, still falls short of an adequate scientific method for estimating true infection prevalence.

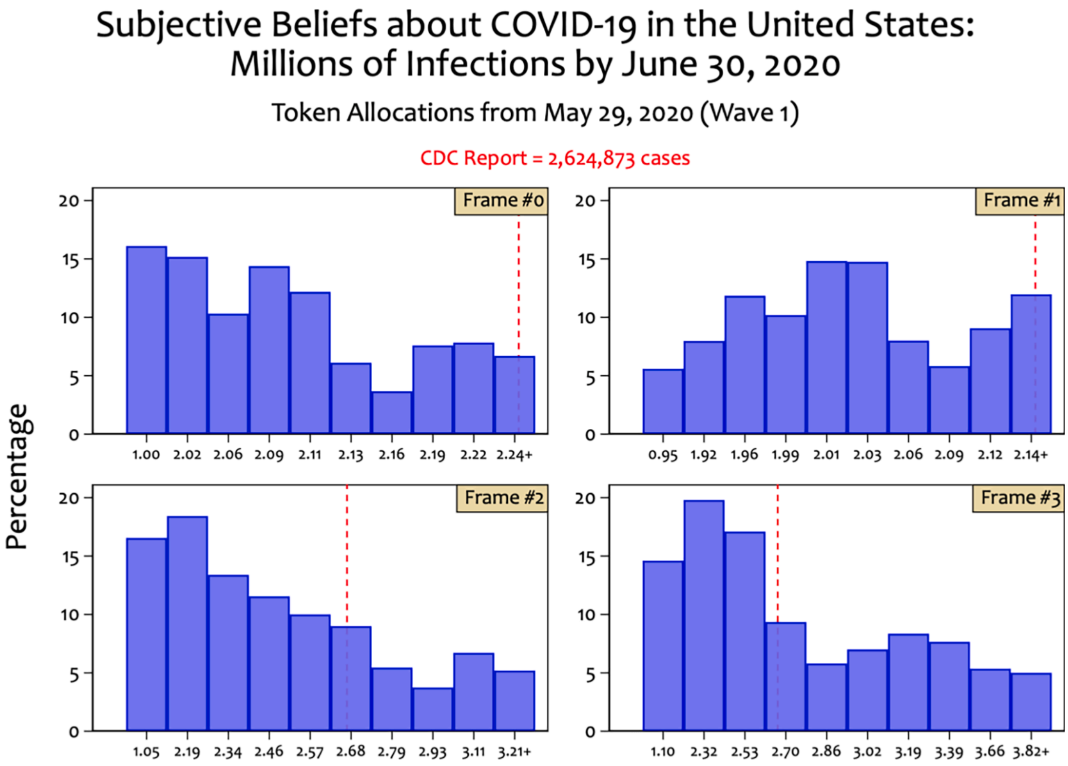


Fig. 4. Beliefs about COVID-19 infections in the U.S. by June 30, 2020.

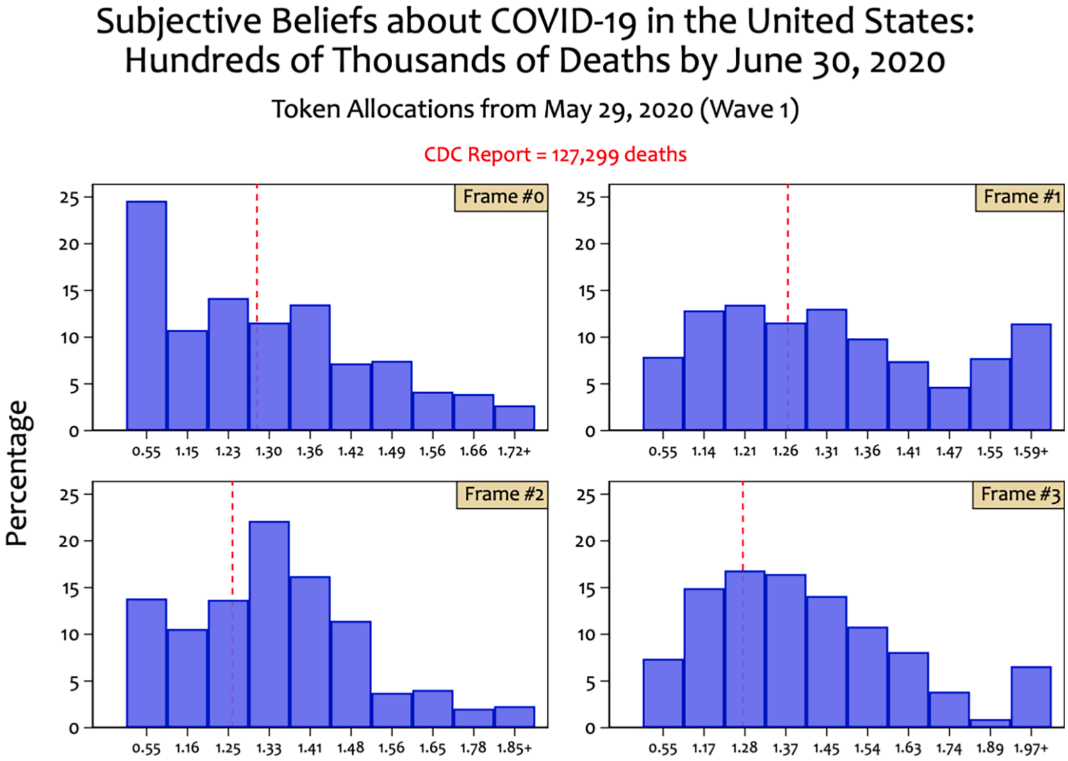


Fig. 5. Beliefs about COVID-19 deaths in the U.S. by June 30, 2020.

space, despite the fact that the IHME considered the most likely level to fall further within the interior region of the event space. By contrast, Fig. 5 shows that the correct answer fell into the "middle" of the event space in every frame. This difference in the accuracy of forecasting infections and deaths is not surprising. Although methods of estimating infections, and reliability of data transmission, vary with the severity of viral spread and the geography of its concentration, deaths are less variable and follow infections with a predictable lag. This is arguably one reason why our subjects' beliefs appear to be more closely calibrated on deaths than on infections.

The potential implications of our research for educational interventions about COVID-19 are clear. While there is no single, well-confirmed consensus theory of health behavior or other-regarding behavior that can ground educational efforts [27], beliefs will play an important role in explaining health behavior, regardless of the specific approach adopted. Beliefs about risk to oneself and to others are fundamental factors in understanding behavior, and are potential levers, therefore, for educational intervention. Meta-analyses show that risk perception has a significant influence on behavior [28], including precautionary reductions in aggregate consumption leading to declines in economic activity [29]. Moreover, the extent to which beliefs influence behaviors, and which beliefs are amenable to educational influence, depends in part on the confidence individuals have in their attitudes. There is also solid evidence that there is heterogeneity across groups in beliefs about health risks [30].

Our sample consists of university students with an average age of 21 years. There is evidence [31] that mortality risks that individuals of a certain age group have current or prior peer experience about are better understood, compared to mortality risks that apply more to older age groups. This finding makes considerable sense, in terms of rational investments in knowledge of mortality risks. However, it suggests that the beliefs of younger adults might not be well adjusted early in the pandemic, when the vast majority of mortalities occurred among (much) older adults.

Our study provides rich data about beliefs and related factors that the literature suggests are necessary to ground educational interventions. Because we elicit incentivized measures of beliefs, as well as the spread of confidence in various COVID-19 outcomes, we have fine-grained detail that is seldom available in educational interventions. Individuals who have very focused beliefs, and discount alternative outcomes strongly, will respond to information differently than individuals who give more credibility to alternative degrees of COVID-19 risk. In Bayesian statistical terms, those with more diffuse priors should respond more strongly to new information than those with tighter priors. We will also be able to investigate how our participant's beliefs about COVID-19 vary according to demographic characteristics, the primary sources of news subjects used to inform themselves about the course of the pandemic, and incentivized elicitations of risk attitudes and time discounting, which we also included in the study. These additional variables could allow one to target public health and educational interventions to particular groups on the basis of their beliefs, and the extent to which they are more or less receptive to new information about risks posed by the virus and attendant mitigation measures.

Educational interventions around COVID-19 using only hypothetical survey information about beliefs, and no evidence about how confidently those beliefs are held, are likely to be unproductive. The methods presented here offer a more useful guide for getting the evidence needed to design successful educational interventions.

5. Conclusion

We conducted an incentivized, experimental study on the beliefs of individuals about COVID-19 prevalence and mortality with six temporally-spaced waves between May and November, 2020. Our experimental design allows us to draw direct, simple inferences about whether those beliefs differ from publicly salient epidemiological models of infections and deaths due to COVID-19. We find that the beliefs of individuals about both infections and deaths differ markedly from epidemiologically informed models. Our study has implications for the dissemination of scientific information, and could be used to tailor public health and educational interventions to people most receptive to risk mitigation efforts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported here.

Supplementary materials

Supplementary materials for this article can be found online at https://doi.org/10.1016/j.ymeth.2021.04.003. The data and code for this article are at https://github.com/andrehofmeyr/Methods-SI.

References

[1] G.W. Harrison, A. Hofmeyr, H. Kincaid, B. Monroe, D. Ross, M. Schneider, J. T. Swarthout. Subjective beliefs and economic preferences during the COVID-19 pandemic. Working Paper 2020-22, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2020.
[2] N.P. Jewell, J.A. Lewnard, B.L. Jewell, Caution warranted: using the Institute for Health Metrics and Evaluation model for predicting the course of the COVID-19 pandemic, Ann. Intern. Med. 173 (3) (2020) 226–227.
[3] R. Marchant, N.I. Samia, O. Rosen, M.A. Tanner, S. Cripps, Learning as we go: an examination of the statistical accuracy of COVID19 daily death count predictions, medRxiv (2020), 20062257.
[4] Piper K., This coronavirus model keeps being wrong. Why are we still listening to it? 2020. Accessed: March 30, 2021. Available from: https://www.vox.com/future-perfect/2020/5/2/21241261/coronavirus-modeling-us-deaths-ihme-pandemic.
[5] F. Galton, Vox Populi, Nature 75 (1949) (1907) 450–451.
[6] F.T. Juster, Consumer buying intentions and purchase probability: an experiment in survey design, J. Am. Stat. Assoc. 61 (315) (1966) 658–696.
[7] A. Delavande, X. Giné, D. McKenzie, Measuring subjective expectations in developing countries: a critical review and new evidence, J. Dev. Econ. 94 (2) (2011) 151–163.
[8] C.F. Manski, Measuring expectations, Econometrica. 72 (5) (2004) 1329–1376.
[9] G.W. Harrison. Hypothetical surveys or incentivized scoring rules for eliciting subjective belief distributions? Working Paper 2014-05, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2014.
[10] T. Fetzer, L. Hensel, J. Hermle, C. Roth, Coronavirus perceptions and economic anxiety, Forthcom. Rev. Econ. Statist. (2021).
[11] L.J. Savage. The foundations of statistics (second edition), Dover Publications, New York, NY, 1972.
[12] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3.
[13] L.J. Savage, Elicitation of personal probabilities and expectations, J. Am. Stat. Assoc. 66 (336) (1971) 783–801.
[14] J.E. Matheson, R.L. Winkler, Scoring rules for continuous probability distributions, Manage. Sci. 22 (10) (1976) 1087–1096.
[15] G.W. Harrison, J. Martínez-Correa, J.T. Swarthout, E.R. Ulm, Scoring rules for subjective probability distributions, J. Econ. Behav. Organ. 134 (2017) 430–448.
[16] D.J. Danker, M.M. Luecke, Background on FOMC meeting minutes, Fed. Reser. Bull. (2005) 175–179.
[17] G.W. Harrison, R.D. Phillips, Subjective beliefs and the statistical forecasts of financial risks: the Chief Risk Officer project, in: T.J. Andersen (Ed.), Contemporary Challenges in Risk Management, Palgrave Macmillan, New York, NY, 2014.
[18] A. Di Girolamo, G.W. Harrison, M.I. Lau, J.T. Swarthout, Subjective belief distributions and the characterization of economic literacy, J. Behav. Exp. Econ. 59 (2015) 1–12.
[19] J.K. Kruschke, Rejecting or accepting parameter values in Bayesian estimation, Adv. Methods Pract. Psychol. Sci. 1 (2) (2018) 270–280.
[20] G.W. Harrison, J.T. Swarthout. Belief distributions, Bayes rule and Bayesian overconfidence. Working Paper 2020-11, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2020.
[21] C. Avery, W. Bossert, A. Clark, G. Ellison, S.F. Ellison, An economist's guide to epidemiology models of infectious disease, J. Econ. Perspectives 34 (4) (2020) 79–104.
[22] Cochrane J.H., The grumpy economist [Internet]. 2020. Accessed: March 22, 2021. Available from: https://johnhcochrane.blogspot.com/2020/05/an-sir-model-with-behavior.html.
[23] Gu Y., Evaluation of COVID-19 models. 2021. Accessed: March 22, 2021. Available from: https://github.com/youyanggu/covid19-forecast-hub-evaluation.
[24] B. Xu, M.U.G. Kraemer, B. Gutierrez, S. Mekaru, K. Sewalk, A. Loskill, L. Wang, E. Cohn, S. Hill, A. Zarebski, S. Li, C.-H. Wu, E. Hulland, J. Morgan, S. Scarpino, J. Brownstein, O. Pybus, D. Pigott, M. Kraemer, Open access epidemiological data

from the COVID-19 outbreak, Lancet Infect. Dis. 20 (5) (2020) 534, https://doi.org/10.1016/S1473-3099(20)30119-5.

[25] T.P.H. Lin, K.H. Wan, S.S. Huang, J.B. Jonas, D.S.C. Hui, D.S.C. Lam, Death tolls of COVID-19: where come the fallacies and ways to make them more accurate, Global Public Health. 15 (10) (2020) 1582–1587.

[26] D. Benatia, R. Godefroy, J. Lewis, Estimating COVID-19 prevalence in the United States: a sample selection model approach, Lancet (2020), https://doi.org/10.2139/ssrn.3578760.

[27] K. Glanz, D.B. Bishop, The role of behavioral science theory in development and implementation of public health interventions, Annu. Rev. Public Health 31 (1) (2010) 399–418.

[28] P. Sheeran, P.R. Harris, T. Epton, Does heightening risk appraisals change people's intentions and behavior? a meta-analysis of experimental studies, Psychol. Bull. 140 (2) (2014) 511–543.

[29] A. Goolsbee, C. Syverson, Fear, lockdown, and diversion: comparing drivers of pandemic economic decline 2020, J. Public Econ. 193 (2021), https://doi.org/10.1016/j.jpubeco.2020.104311, 104311.

[30] P.D. Lunn, C.A. Belton, C. Lavin, F.P. McGowan, S. Timmons, D.A. Robertson, Using behavioral science to help fight the coronavirus, J. Behav. Public Admin. 3 (2020), https://doi.org/10.30636/jbpa.31.147.

[31] G.W. Harrison, E.E. Rutström, Eliciting subjective beliefs about mortality risk orderings, Environ. Resour. Econ. 33 (3) (2006) 325–346.