Multiple Forecast Visualizations (MFVs): Trade-offs in Trust and Performance in Multiple COVID-19 Forecast Visualizations

Lace Padilla, Racquel Fygenson, Spencer C. Castro, Enrico Bertini

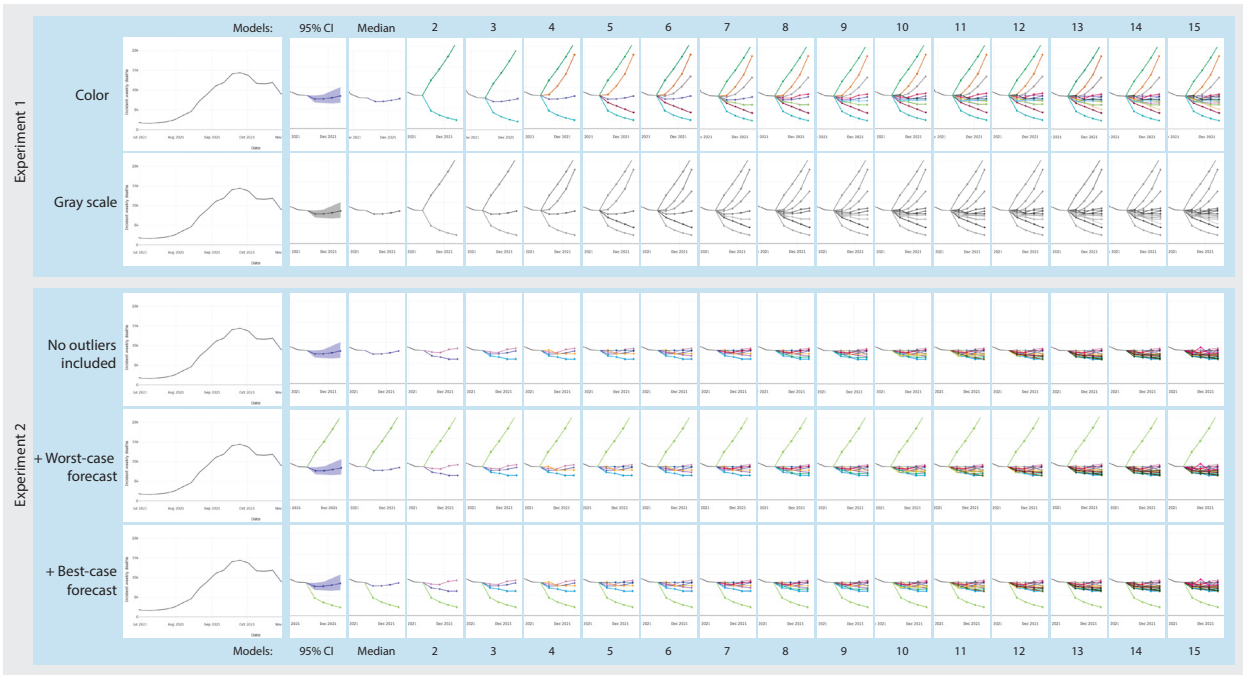

Fig. 1. Stimuli used in Experiments 1 and 2 showing COVID-19 mortality forecasts for November 13, 2021 in the US. Each line depicts a different group's forecast, and the experiments examined the impact of the number of forecasts shown on trust and predictions of the COVID-19 trends. Each participant was shown the 16 stimuli in one row of this figure in a randomized order.

Abstract— The prevalence of inadequate SARS-COV-2 (COVID-19) responses may indicate a lack of trust in forecasts and risk communication. However, no work has empirically tested how multiple forecast visualization choices impact trust and task-based performance. The three studies presented in this paper (N = 1299) examine how visualization choices impact trust in COVID-19 mortality forecasts and how they influence performance in a trend prediction task. These studies focus on line charts populated with real-time COVID-19 data that varied the number and color encoding of the forecasts and the presence of best/worst-case forecasts. The studies reveal that trust in COVID-19 forecast visualizations initially increases with the number of forecasts and then plateaus after 6-9 forecasts. However, participants were most trusting of visualizations that showed less visual information, including a 95% confidence interval, single forecast, and grayscale encoded forecasts. Participants maintained high trust in intervals labeled with 50% and 25% and did not proportionally scale their trust to the indicated interval size. Despite the high trust, the 95% CI condition was the most likely to evoke predictions that did not correspond with the actual COVID-19 trend. Qualitative analysis of participants' strategies confirmed that many participants trusted both the simplistic visualizations and those with numerous forecasts. This work provides practical guides for how COVID-19 forecast visualizations influence trust, including recommendations for identifying the range where forecasts balance trade-offs between trust and task-based performance.

Index Terms—COVID-19, multiple forecast visualizations, uncertainty visualization, line charts, time-series data

1 Introduction

- *L. Padilla and C. Castro are with the University of California Merced, Email: lace.padilla@ucmerced.edu, scastro39@ucmerced.edu*
- *R. Fygenson is with New York University, E-mail: racquel.fygenson@nyu.edu racquel.fygenson@nyu.edu*
- *E. Bertini is with Northeastern University, E-mail: e.bertini@northeastern.edu*

Manuscript received 31 March 2022; revised 1 July 2022; accepted 8 August 2022. Date of publication 27 September 2022; date of current version 2 December 2022. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2022.3209457, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2022.3209457

Ensuring trust during public health emergencies is a critical first step in encouraging the public to follow health officials' recommendations. For example, the SARS-CoV-2 (also known as COVID-19) pandemic demonstrated that reduced trust in public health mandates results in less compliance with orders [3]. Researchers have extensively studied factors that contribute to trust in health care communication (e.g., [3, 36, 63]), and proposed several domain-dependent trust definitions [2, 16, 62]. Whereas prior work on COVID-19 and trust uses more general definitions (e.g., [3,13,36,60,61]), we use Mayr et al.'s [44] definition of trust in information visualizations: "trust is the user's implicit or explicit tendency to rely on a visualization and to build on the information displayed." Visualization techniques used to communicate COVID-19 forecasts are an understudied factor that contributes to trust. As millions

worldwide turn to visualizations to understand their pandemic health risks [37], we need to understand how visualization decisions impact trust, ultimately influencing willingness to follow recommendations.

This work tested if and how varying visualization techniques impact trust in the forecast shown when depicting current COVID-19 mortality data. As a preview, we found evidence for a trade-off effect where uncertainty communication optimizes trust and performance in a trend prediction task. We find that lower complexity visualizations (e.g., 95% confidence intervals and point-estimate forecasts), although highly trusted, produce poor decision quality as they fail to depict important nuances in the forecasts. On the other hand, providing too much uncertainty information (in this case, more than nine forecast models) provides no additional judgment benefits and can adversely impact the trust of some viewers. We also found that viewers' trust in confidence intervals can be misplaced and disproportionately scaled.

This work's primary contribution is to present empirical evidence for the link between COVID-19 forecast visualization design choices and trust. The following experiments provide several empirical and quantitative analyses of how people conceptualize COVID-19 forecasts and factors that contribute to modeling uncertainty. We offer visualization recommendations to optimize trust and performance in a trend estimation task with pandemic visualizations. The insights from this work can help risk communicators make informed decisions about the trade-offs between trust and task-based performance in pandemic forecasts, which could improve health and public safety in future pandemics and other longitudinal hazards such as climate change.

2 Related Work: Uncertainty Visualization

When depicting uncertainty, visualization designers choose between two families of methods (for a review of methods, see [47]). The first group consists of summary annotations (e.g., confidence intervals and mean plots). Although popular–60% of COVID-19 line charts examined in an extensive review used confidence intervals [66]–decades of research have demonstrated that viewers misinterpret these annotations [5, 8, 10, 19, 21, 27, 28, 31, 49, 57, 58], despite targeted training [6] and regardless of their level of expertise [5].

The second family of methods consists of distributional visualizations (e.g., gradient, violin, quantile dot plots, and ensemble line charts). These visualizations often result in increased reader performance over summary annotations [8, 10, 14, 24, 31–33, 49, 58], text-based explanations of distributions, and visualizations without any uncertainty [8, 14]. The family of distributional visualizations can be split further into explicit and implicit encodings of uncertainty [47]. Explicit uncertainty visualizations, like quantile dot plots [33] and density plots, encode confidence or probability via marks. Implicit uncertainty visualizations use coinciding depictions of multiple possible outcomes to communicate uncertainty without quantification of a concrete metric [12] (e.g., ensemble charts [41] and hypothetical outcome plots [25]).

This study examines *multiple forecast visualizations* (MFVs), a type of chart where uncertainty is encoded implicitly through the disagreement (or lack thereof) of multiple forecasts plotted in the same space, which informs readers about the range, shape, and concentration of predictions [46]. This technique depicts many forecasts at once, but instead of sampling those forecasts from a distribution–as is done in ensemble plots [38, 40, 41] or hypothetical outcome plots [25]–an MFV depicts several unique predictions from different forecasting entities. There is currently no guidance for intelligently selecting forecasts in MFVs. One could include all available forecasts, but there could be dozens, resulting in overplotting. The following experiments examine the trade-offs between trust, intelligibility, and performance in a trend prediction task to provide actionable guidance for MFV design.

2.1 Evaluations of COVID-19 Visualizations

In a review of 668 COVID-19 visualizations, Zhang et al. found that, of those that showed forecasts with uncertainty, the two most common approaches were those that used confidence intervals (60%) or multiple models or scenarios to express uncertainty (29%; [66]). Our prior work comparing COVID-19 forecast visualizations depicting intervals, no uncertainty (point-estimate forecast plots), and MFVs found that MFVs were most likely to change participants' beliefs about the COVID-19 risk to themselves and others [46]. In the aforementioned work, consistently impactful visualizations presented all possible forecast models (35) or 6 forecast models selected by a modeling expert. These two approaches led people to believe that they and others were at more risk from COVID-19 than before viewing the forecasts, and motivated the current investigation into trade-offs between trust and performance.

Other researchers have compared COVID-19 time-series forecast visualizations with no uncertainty (point-estimate plots), 95% confidence intervals, and ensemble plots [38]. Leffrang et al. asked participants to estimate the number of hospitalizations one, two, and three weeks into the future before and after viewing a visualization. They found that participants who viewed the ensembles were less willing to update their hospitalization estimates to align with the forecasts compared to point-estimates and 95% CIs [38].

Researchers have also considered the axis used in COVID-19 visualizations and have found that a cumulative y-axis scale (e.g., additive counts) of COVID-19 data leads to stable risk interpretations [46,56]. In contrast, these studies found that an incident scale (e.g., counts summed over short time intervals such as a week) produces more variable [46] and riskier [56] interpretations of the COVID-19 data.

2.2 Trust

Recent research suggests that the public's trust in COVID-19 information correlates with the likelihood of taking preventative measures to stymie the pandemic [43, 45]. Trust in information visualization is multifaceted [44], and a recent study has suggested that it is closely tied to the perceived transparency of the communication [22], which increases with quantity and perceived quality of information [59].

On the other hand, increased transparency can also lead to audience speculation, mistrust of presented data, and even mistrust of entities producing reports of data [9, 18]. The beneficial and detrimental impacts of increasing information availability parallel existing visualization research. Despite some visualization studies suggesting that more information availability through visualizing uncertainty can increase trust [26, 29, 34], complementary studies caution that too much visual information can increase audience confusion, negatively impacting the accuracy of interpretations [39, 50]. Unfortunately, only few studies have investigated trust within the context of a pandemic-specific visualizations [1, 38].

Our studies provide an examination of the relationship between perceived trust and the number of pandemic forecasts shown, the shape of the pandemic forecasts, and coloring vs grayscale designs. In doing so, we explore the relationship between users' self-reported general trust and the balance of disclosure versus performance.

3 Methods and Aims

Our current work examines the effects of three visualization design choices for MFVs on trust (number of forecasts, color, and best/worst-case forecasts). First, we examine the impact of the number of forecasts visualized on trust. Although prior work found that MFVs produced the largest changes in COVID-19 risk estimates [46], the optimal number of forecasts is unclear. Our preregistered hypothesis stated that (**H1**) the relationship between trust and the number of forecasts shown would be positive (**H1A**) and nonlinear such that there would be a trust asymptote (**H1B**). This hypothesis aligns with the theory that trust estimates incorporate thoroughness, disclosure, and clarity [64]. Showing numerous forecasts increases disclosure and thoroughness but may reduce clarity.

The second design choice we examined entails the impact of color on trust estimates. Our preregistered hypothesis stated that (**H2**) participants viewing forecasts with different colors would perceive the forecasts as less trustworthy than those in grayscale (**H2A**), increasing this difference as the number of forecasts increases (i.e., an interaction; **H2B**). We predict that adding color to each of the forecasts may add superfluous complexity and reduce clarity in the case of the COVID-19 MFVs with numerous forecasts.

Finally, we examined the design choice to present or omit best- and worst-case model forecasts (Experiment 2). We were inspired to examine the impact of best- and worst-case model forecasts by

stakeholders from a government agency who indicated that it was vital to show the worst-case scenarios for planning risk-reduction policies. Our preregistered hypothesis stated that (**H3**) best- and worst-case model forecasts will lead to less trust (**H3A**), especially when fewer forecasts are shown (**H3B**). We predicted that visualizations that include an extreme forecast will draw viewers' attention to overall forecast disagreement and, therefore, be deemed less trustworthy. Further, we predicted that the negative impact of extreme forecasts would be reduced when more forecasts were shown because more models would minimize the distrust stoked by a conflicting extreme forecast (i.e., interaction between forecast number and presence of outliers; **H3B**).

To test the impact of these design choices on trust, we conducted two online experiments in November 2021 with real-time COVID-19 mortality data from the United States. Experiment 1 examined the impact of the number of forecasts and the presence or absence of color on trust estimates. Experiment 2 examined the impact of the number of forecasts and best- and worst-case forecasts on trust. Both experiments used a median ensemble forecast that showed no uncertainty and a 95% confidence interval forecast as controls. We conducted a third follow-up experiment in February of 2022 to further understand the effects of the 95% confidence interval label on trust estimates.

3.1 Design

We preregistered the first two experiments on the Open Science Framework (`https://osf.io/e2fnd`). For Experiment 1, we used a mixed design where the online survey software Qualtrics [53] randomly assigned participants to a color or grayscale group. Within-subjects manipulations included the number of forecasts shown (1-15) and a 95% confidence interval (see, Figure 1 top two rows). Participants completed 16 trials total (one trial per condition) in a randomized order.

Experiment 2 also used a mixed design where participants were randomly assigned to one of three groups: 1) forecasts that included a best-case scenario, 2) forecasts that included a worst-case scenario, and 3) forecasts that included neither best nor worst-case scenarios, referred to as the base forecasts (see, Figure 1 bottom three rows). The within-subjects manipulations were the same as in Experiment 1; 16 trials in a randomized order (forecasts 1-15 and a CI 95).

We conducted Experiment 3 as a follow-up study to examine whether the 95% confidence interval caption impacted trust ratings. Experiment 3 was a full between-subject experiment where participants were randomly assigned to one of eight groups; they saw the same 95% confidence interval forecast but with different figure captions and labels that we manipulated. In four groups, participants viewed a 95% confidence interval forecast with text captions indicating that the interval was 99%, 95%, 50%, or 25%. The other four groups viewed the same figure captions, but additional annotations reiterated the confidence interval size. Examples of additional annotations indicating 25% are shown in Figure 2. The additional annotations were tested to ensure that participants noticed the confidence interval manipulation.

3.2 Stimuli

The visualizations were generated using the Reich Lab's COVID-19 Forecast Hub online visualization application [55] using Plotly [51]. The Reich Lab's COVID-19 Forecast Hub is a central repository of COVID-19 forecasts and predictions from over 50 international research groups and provides a visualization tool to display the forecasts [11]. The COVID-19 Forecast Hub produces a weighted median forecast at horizons of 1 through 4 weeks ahead for 10 forecasts with the best performance in the 12 previous weeks. They generate a 95% confidence interval from this ensemble median.

We selected the forecasts for the stimuli in Experiment 1 to represent the largest range of predictions, including the most optimistic and pessimistic forecasts (as seen in Figure 1, top two rows). We tested charts with more forecasts than people would likely be able to discriminate (given the line thickness and spread) because we thought they might be sensitive to the holistic information in the chart. We then added additional forecasts to fill out the distribution while attempting to convey the most likely trend in the mortality data, as described by the median ensemble forecast. We also attempted to reduce overplotting where possible. The 15 total forecasts were selected to reflect the upper bounds of forecast numbers that we could distinguish for the spread of forecasts in Experiment 1. We determined that other design components of charts in the COVID-19 Forecast Hub (e.g., sans-serif font, common date formatting, neutral color scheme) were simple, legible, and unlikely to bias results, and thus maintained them in all stimuli. Grayscale versions of these visualizations were created in Adobe Photoshop.

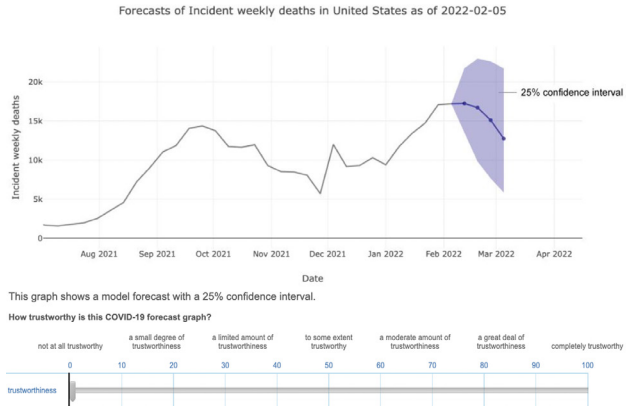

Fig. 2. Example 95% confidence interval from Experiment 3 with a caption and direct annotation reading 25% confidence interval and the trust rating scale.

For Experiment 2, we selected forecasts that approximated the range of the 95% confidence intervals (Figure 1, third row). The differences in the spread of the MVFs and CI95 in Experiment 1 could be a confound, and we sought to control for this in Experiment 2. Additionally, we tested the impact of worst- and best-case forecasts by adding them to each visualization, including the CI95 (Figure 1, bottom two rows). Note that due to the smaller range used in Experiment 2, which matched the range of the CI95, more overplotting occurred. We presented the stimuli in Experiments 1 and 2 at 905 x 437 px to roughly preserve the charts dimensions as they appeared in the COVID-19 Forecast Hub.

For Experiment 3, the CI95 was the only visualization tested. Each group viewed the CI95 visualization with different standard captions (25%, 50%, 95%, and 99%) or additional captions with direct annotations, resulting in eight groups. The direct annotations were added to the forecasts in Adobe Illustrator (see Figure 2). The images were presented online at 1000 x 467px. We subjectively confirmed stimuli the experiments were crisp and legible at the resolution of 72 PPI.

3.3 Procedure and Tasks

Participants completed an IRB consent protocol, then read the following instructions: *"**Instructions:** In this experiment, you will see different COVID-19 mortality forecast graphs for the United States like the one below. Your job is to rate how trustworthy you think the forecast is on a scale from 0 (Not at all Trustworthy) - 100 (Completely Trustworthy), like the one below. Please try to use the entire scale. You will also be asked a follow-up question about each graph.*

***Forecasts:** Each line with dots shows actual COVID-19 mortality forecasts for the next two weeks in the US. Each line shows one model that different groups of researchers created.*

***Task:** We would like to know how trustworthy you think the graph is as a whole, taking into account any forecasts shown. Knowing how trustworthy you think the graph is will help us guide researchers about which forecasts to show to the public."*

For our trust rating scale, participants dragged a slider (0-100, shown in Figure 2) with empirically validated textual anchors (i.e., prior work demonstrates that the anchors are perceived as equally distant [7]) to indicate their trust. We selected this trust measure based on prior work that used such self-report ratings to examine participants' trust in the context of COVID-19 [3, 13, 36, 60, 61]. Although self-report allows participants to define trust, we felt that varied individual definitions maintained ecological validity, despite reducing experimental control.

Further, for the within-subjects trials, we determined changes in participants' individual understanding of trust were due to the number of MFVs or the intervals (which were the within-subjects conditions). We describe the limitations of this approach in Section 5.1. To understand more about participants' definitions of trust, we also asked open-ended questions at the end of the experiment. Finally, prior work suggests that people do not perceive trust and distrust as perfectly inversely correlated [35, 42], and therefore, our measure asking participants to indicate *trustworthiness* does not evaluate distrust.

Following the trust ratings, participants answered the question, *"Based on this forecast, do you think the COVID-19 deaths in the US over the next two weeks will: Increase, Decrease, Stay the Same, Unsure?"* After the 16 primary trials, participants completed a manipulation check where they counted the number of forecasts they viewed in a subset of the stimuli. The goal of this manipulation check was to determine if viewers could discriminate each forecast. At the end of the experiment, participants provided demographic information and open-ended responses to the following: *"Please describe in as much detail as you can how you made your trust judgments."*, *"Why do you think there is uncertainty in COVID-19 forecasts?"*, and *"Why do you think the various forecasts made different predictions about the COVID-19 deaths in the US?."* The procedure and tasks for Experiment 3 were the same as in the prior experiments, but participants completed only one trust judgment about the confidence interval visualization and only answered one open-ended question about their strategy in the task (Q1).

3.4 Participants

The experiments were conducted online with populations of participants from Prolific [52]. We compensated participants $2.30 to partake in the studies, which took roughly 10 minutes to complete. The prescreening criteria dictated that participants were over 18 years old, were currently living in the United States, and were not allowed to participate in more than one of the experiments in this report. We selected the preregistered power analysis parameters based on prior work on decision-making with hurricane forecast visualizations that had a small-medium effect size (Cohen's $d = 0.29$) [49]. Because of the difference between Padilla et al. [49] and the current study, we reduced the prediction to a small effect to be conservative (Cohen's $f^2 = .11$). One hundred people participated in each group in each experiment (see the preregistered report for details).

In Experiment 1, participants were 200 US residents (52% Women, *Mean* age = 24.72 years, *SD* = 6.56 years). In Experiment 2, participants were 299 US residents (52.51% Women, *Mean* age = 25.81 years, *SD* = 9.26 years). We collected one fewer participant in the base forecast group because Prolific allowed one individual to receive credit without participating. In Experiment 3, participants were 800 US residents (51% Women, *Mean* age = 37.5 years, *SD* = 14.29 years).

4 Results

We conducted a quantitative analysis of the trust ratings and COVID-19 trend estimates, qualitative analyses of the open-ended questions, and a manipulation check. We preregistered the trust analysis in the first two studies, which utilized linear mixed-effects models from the lme4 package [4] to fit the data in the statistical computing and visualization environment R [54]. This analysis determined the relationship between trust and the visualization conditions. In the qualitative analysis, two coders read the 1500 responses from Experiments 1 and 2, and then indicated their interpretations of the participants' self-reported strategies/beliefs about COVID-19 forecasts. Finally, we reported a manipulation check of how many forecasts participants could distinguish in the visualizations. We conducted the same sequence of analyses for the first two experiments and focused on the quantitative analysis of trust for Experiment 3.

This section presents statistics from ANOVA and regression analyses. The ANOVA analysis utilizes the chi-squared (χ^2) goodness-of-fit to determine the most parsimonious regression model. For this test, $p < .05$ provides evidence that the most complex model in the comparison has the best fit. In the regression analysis, an example of a β interpretation is as follows: for every additional forecast model added, trust will increase by β, while holding all other variables constant. The t-value (t) is the β divided by the standard error of that variable. In all analyses, the number in parentheses that follows a test statistic is the degrees of freedom. This reporting style adheres to the American Psychological Association statistics reporting style guide [30].

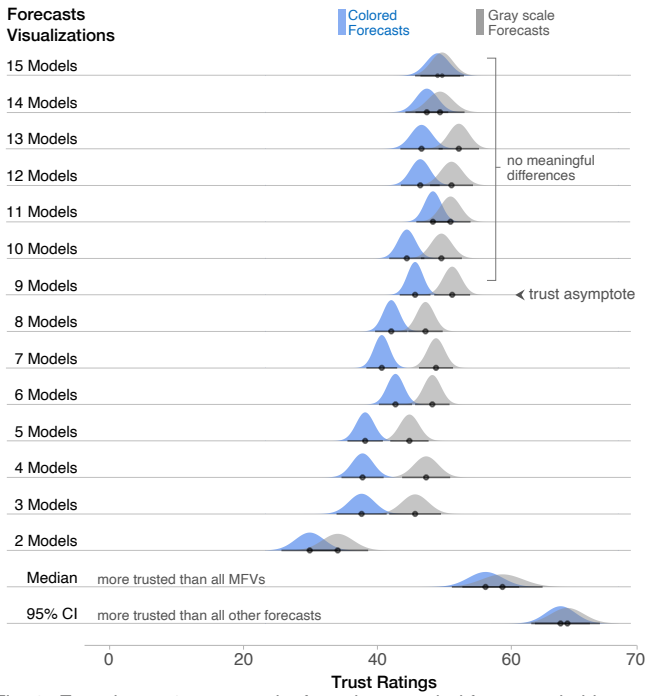

Fig. 3. Experiment 1 trust results for color encoded forecasts in blue and grayscale in gray.

4.1 Experiment 1

4.1.1 Quantitative Trust Results

In the first experiment, we tested the impact of 1) the number of forecasts visualized and 2) color-encoded forecasts on participants' trust ratings. Complete data, analysis code, and model outputs can be viewed in the supplemental materials (OSF link).

Omnibus Analysis. We predicted that increasing the number of forecasts would increase trust (**H1A**) but stop doing so at some point (**H1B**), that color would reduce trust due to an increase in visual complexity which would reduce clarity, when compared to grayscale (**H2A**), and that increasing forecasts would enhance color's effect (i.e., an interaction; **H2B**). To test these hypotheses, we specified multiple linear-mixed effects models and compared these models via the ANOVA function from the *car* package in R [15]. To predict variance in trust ratings, we compared the most parsimonious model, which specified the number of forecast visualizations (i.e., Forecast 1-15 and CI 95; **H1A**) and color (i.e., color vs grayscale) as fixed effects, to simpler models ($\chi^2(1)$=4.63, p =0.03), and to a model that included an interaction between $ForecastVis * Color$, which did not significantly improve model fit ($\chi^2(15)$=16.22, $p = 0.37$). All models specified random intercepts for each participant. The R-programming software notation for the winning model is: $Trust \sim ForecastVis + Color + (1|Participant)$. We detail the R notation in this paper to save space rather than including the full model equation. As model comparisons selected the equation with no interaction term, the analysis suggests that the relationship between color and grayscale does not significantly vary across the forecast visualizations (**H2B** not accepted; see Figure 3).

Forecast Type Analysis. For analyses of main effects, we first compared all levels of fixed effects to the median ensemble and grayscale forecast as the referents. We did not preregister predictions for the 95% CI and included it as a control condition in this study. Participants rated the median ensemble forecast as more trusted than all other MFVs but less trusted than the 95% CI (the model output is shown in Table 1 row B). We also conducted the same analysis but with the

95% CI as the referent, confirming that participants trusted the 95% CI significantly more than all other visualizations (Table 1 row A). These findings demonstrate that people place the most trust in visualizations with less visual information, the point-based estimate and the 95% CI (see Leffrang and Müller [38] for similar findings).

Color Encoded Analysis. The other main effect in the original model (Section 4.1.1, Omnibus Analysis) determined that when trust was averaged across all of the forecast visualization types, participants trusted the grayscale encoded visualizations ($Mean$ = 49.8, SD = 23.9) significantly more than the visualizations encoded with color ($Mean$ = 45, SD = 24.5, β = -4.8, $t(200)$ = -2.16, p = .031, CI[-9.2, -.4]). On average, grayscale visualizations were 4.8% more trusted than color-encoded visualizations. This result supports the first part of our hypothesis that color-encodings would decrease trust (**H2A**). However, we did not find support that the relationship between the number of forecasts and trust would be different for color vs grayscale encodings (**H2B** not accepted). We found the main effect only where grayscale encodings were more trusted overall. However, because this main effect is collapsed across the forecast visualization types, it does not reveal if some forecasts lack a meaningful difference in trust between grayscale and color. The absence of a significant interaction suggests no evidence of a difference in the effect of color across the forecast types. Nevertheless, a visual analysis of Figure 3 indicates less of an effect of color for the median, CI95, 14, and 15 forecasts.

Asymptote Analysis. We hypothesized that viewers would associate more forecasts with increased transparency and thoroughness (**H1A**), but, if too many forecasts were shown, clarity would decrease. Thus, we predicted that the relationship between trust and the number of forecasts would be positive (**H1A**) but nonlinear—possibly showing asymptotic growth (**H1B**). Narrowing in on the visualizations that conveyed uncertainty indirectly (2-15 forecasts), we investigated if at any point increasing the number of forecasts stopped improving reported trust. To do so, we conducted the same analysis as before (Section 4.1.1, Omnibus Analysis) but with a sequence of equations where we specified 2-10 forecasts as the referent group. Comparing the forecasts to one another by systematically changing the referent allowed us to identify a trust asymptote, which we defined as the point where there were no longer meaningful differences in trust. As seen in Figure 3, participants trusted the visualization of 9 forecasts more than those with 2-5, 7, and 8 forecasts (**H1A** partly accepted), but there was no meaningful difference in trust for stimuli with 6 and 9-15 forecasts (**H1B** accepted; Table 1 row C). This finding suggests that trust generally increases as the number of forecasts increases from 2 to 9, after which it plateaus. Note that we did not test more than 15 forecasts, so, although not indicated in our study, there could be a point at which increasing the number of forecasts shown *decreases* trust. These findings support our hypothesis that trust increases with the number of forecasts shown in a nonlinear fashion (**H1** accepted).

4.1.2 COVID-19 Trend Prediction

Participants indicated their belief about the change in the rate of COVID-19 deaths in the US over the next two weeks (e.g., increasing, decreasing, staying the same, or unsure). The number of mortalities during the week of the forecast (Nov. 13th, 2021) in the US was 8457, which declined 32.4% in two weeks (Nov. 27th 2021, 5716). We focus on the prediction of "decreasing" as being consistent with the subsequent mortality trend. As the primary goal of forecast visualizations is to help viewers accurately predict the forecasted event, we felt that interpreting the results in relation to the actual COVID-19 trend two weeks following the forecast was an objective benchmark. We did not preregister hypotheses regarding trend predictions because we did not know what the COVID-19 trends would be two weeks in the future.

We compared multiple linear-mixed effects equations to examine whether the forecast visualizations impacted participants' interpretations of the COVID-19 trends. The most parsimonious equation used the forecast visualizations (Forecast 1-15 and CI 95) and color (color vs grayscale) as fixed effects to predict variance in trend judgments, with random intercepts for each participant. We coded participants' responses as -1 (decreasing), 0 (staying the same), and 1 (increasing).

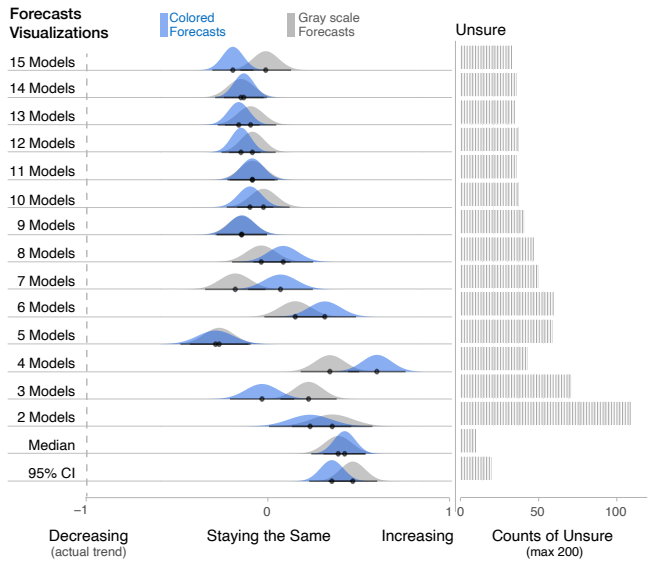

Fig. 4. Trend analysis for Experiment 1 where -1 is decreasing (actual trend), and where color encoded forecasts are in blue and grayscale in gray. Right of the figure shows counts of "unsure" responses.

We excluded the unsure responses because they did not have a clear order in our coding system, but we report the number of people who responded with unsure in Figure 4. The R notation for this model was $JudgmentCode \sim ForecastVis + Color + (1|Participant)$. We specified the 95% CI and grayscale encoded forecasts as the referents.

As seen in Figure 4, we found no main effect of color (β = -.01, $t(196)$ = -.26, p = .80), CI[-.12, .09]). The 95% CI and median ensemble forecasts were significantly less likely to lead to predictions of the actual COVID-19 trend (i.e., evoking responses that were furthest from -1, i.e., declining) than all the other forecasts except 2- and 4- models (Table 1 row D shows results for CI95). To examine the visualization that was most likely to elicit trend predictions that match the actual outcome, we changed the referent to 5 Forecasts, which revealed greater accuracy than all but two of the other visualizations (Table 1 row E).

When considering the unsure participants (Figure 4, right), many people were unsure when viewing the stimuli with two forecasts. In contrast, only a few participants were unsure when considering the median ensemble and 95% CI forecasts. Visual analysis of the plotted "unsure" counts reveals an inverse correlation between the number of participants who report being unsure and the number of forecasts shown. We also find an inverse correlation between the reported trust of a visualization and the number of participants who are unsure of the visualized trends' directionality.

4.1.3 Manipulation Check

At the end of the experiment, we asked participants to count and report the number of forecast lines in six of the stimuli (charts with 10-15 forecasts). We excluded stimuli with fewer forecasts, assuming nearly 100% accuracy. To evaluate accuracy, any counts within +-1 of the forecasts shown were considered correct. As seen in Table 2, grayscale stimuli averaged 86.83% and color-encoded stimuli averaged 91.83%. In both stimuli groups, participant error increased with the number of forecasts. After conducting a sensitivity analysis in which we excluded participants who failed all manipulation/attention check questions, we found these exclusions did not meaningfully change the results. Therefore, we did not exclude participants from analyses and only report the manipulation check results for thoroughness.

	Exp 1		Exp 2		
MVF	Gray	Color	Base	Best-case	Worst-case
15	70	91	29	29	39
14	91	92	39	32	35
13	83	83	36	45	54
12	89	94	39	49	47
11	93	93	61	60	68
10	95	98	73	66	78
9			80	77	85
8			99	95	91
Mean	86.8	91.8	57.2	56.6	62.1

Table 2. Percent accuracy in the manipulation check where participants counted a subset of forecasts in Experiments 1 and 2.

Experiment	Row	Comparison	Median	2	3	4	5	6	7	8	9	10	11	12	13	14	15
EXP 1 Trust	A	CI 95	**β = -10.5 t = -6.0**	**β = -35.9 t = -20.6**	**β = -26.2 t = -15.1**	**β = -25.3 t = -14.6**	**β =-26.4 t = -15.2**	**β = -22.4 t = -12.9**	**β = -23.2 t = -13.3**	**β = -23.3 t = -13.4**	**β = -19.5 t = -11.20**	**β = -20.9 t = -12.0**	**β = -18.3 t = -10.5**	**β = -19.1 t = -11.0**	**β = -18.5 t = -10.6**	**β = -19.5 t = -11.2**	**β = -18.6 t = -10.7**
	B	Median		**β = -25.4 t = -14.3**	**β = -15.8 t = -9.07**	**β = -14.9 t = -8.56**	**β = -15.9 t = -9.17**	**β = -12 t = -6.88**	**β = -12.7 t = -7.31**	**β = -12.8 t = -7.36**	**β = -9.02 t = -5.18**	**β = -10.4 t = -6.00**	**β = -7.81 t = -4.49**	**β = -8.68 t = -4.99**	**β = -8.04 t = -4.62**	**β = -9.05 t = -5.20**	**β = -8.09 t = -4.65**
	C	9			**β = -6.77 t = -3.89**	**β = -5.87 t = -3.38**	**β = -6.93 t = -3.98**	β = -2.95 t = -1.70	β = -3.70 t = -2.13	β = -3.78 t = -2.17		β = -1.43 t = -0.82	β = 1.21 t = 0.69	β = 0.34 t = 0.19	β = 0.98 t = 0.56	β = -0.04 t = -0.02	β = 0.93 t = 0.53
EXP 1 Trend	D	CI 95	β = .001 t = .01	β = -.12 t = -1.4	**β = -0.31 t = -3.96**	β = 0.05 t = 0.68	**β = -0.68 t = -8.99**	β = -0.19 t = -2.54	**β = -0.45 t = -6.02**	**β = -0.38 t = -5.11**	**β = -0.54 t = -7.42**	**β = -0.45 t = -6.21**	**β = -0.47 t = -6.53**	**β = -0.51 t = -6.97**	**β = -0.52 t = -7.20**	**β = -0.54 t = -7.39**	**β = -0.49 t = -6.84**
	E	5	**β = .7 t = 9.1**	**β = 0.56 t = 6.13**	**β = 0.56 t = 4.55**	**β = 0.37 t = 9.37**		**β = 0.49 t = 6.09**	**β = 0.23 t = 2.95**	**β = 0.30 t = 3.86**	β = 0.14 t = 1.78	**β = 0.23 t = 2.97**	**β = 0.21 t = 2.69**	β = 0.17 t = 2.26	β = 0.16 t = 2.07	β = 0.14 t = 1.88	β = 0.19 t = 2.44
EXP 2 Trust	F	CI 95: Base	**β =-9.08 t = -3.81**	**β = -21.60 t = -9.07**	**β = -13.47 t = -5.66**	**β = -16.91 t = -7.10**	**β = -11.83 t = -4.97**	**β = -13.03 t = -5.47**	**β = -11.67 t = -4.90**	**β = -13.49 t = -5.67**	**β = -13.60 t = -5.71**	**β = -13.29 t = -5.58**	**β = -15.86 t = -6.66**	**β = -15.33 t = -6.44**	**β = -12.74 t = -5.35**	**β = -16.18 t = -6.80**	**β = -19.54 t = -8.20**
	G	CI 95: + Best	**β = -14.75 t = -6.23**	**β = -7.01 t = 2.96**	β = -3.35 t = -1.41	β = -1.48 t = -0.62	β = -0.43 t = -0.18	β = -1.48 t = -0.62	β = 1.30 t = 0.55	β = 1.01 t = 0.43	β = 1.67 t = 0.70	β = 1.44 t = 0.61	β = 3.76 t = 1.59	β = 2.30 t = 0.97	β = 4.83 t = 2.04	β = 1.72 t = 0.73	β = 2.85 t = 1.20
	H	CI 95: + Worst	**β = -16.62 t = 7.01**	**β = -10.92 t = -4.61**	**β = -6.91 t = -2.92**	**β = -8.92 t = -3.76**	**β = -7.74 t = -3.27**	β = -4.49 t = -1.89	β = -3.71 t = -1.57	β = -3.72 t = -1.57	β = -2.07 t = -0.87	β = -2.13 t = -0.90	β = -1.10 t = -0.46	β = -1.57 t = -0.66	β = -0.83 t = -0.35	β = -0.73 t = -0.31	β = -0.89 t = -0.38
EXP 2 Trend	I	CI 95	β = -0.07 t = -1.30	**β = -0.67 t = -12.6**	**β = -0.40 t = -7.99**	**β = -0.58 t = -11.52**	**β = -0.75 t = -15.3**	**β = -0.84 t = -17**	**β = -0.79 t = -16**	**β = -0.86 t = -17.6**	**β = -0.85 t = -17.2**	**β = -0.88 t = -17.8**	**β = -0.86 t = -17.4**	**β = -0.83 t = -16.6**	**β = -0.86 t = -17.2**	**β = -0.86 t = -17.2**	**β = -0.86 t = -16.9**
	J	Median		**β = -0.61 t = -11.1**	**β = -0.33 t = -6.53**	**β = -0.52 t = -10.01**	**β = -0.69 t = -13.7**	**β = -0.77 t = -15.3**	**β = -0.72 t = -14.4**	**β = -0.80 t = -15.9**	**β = -0.78 t = -15.5**	**β = -0.81 t = -16.1**	**β = -0.79 t = -15.7**	**β = -0.76 t = -15.0**	**β = -0.79 t = -15.6**	**β = -0.79 t = -15.6**	**β = -0.79 t = -15.3**

Legend: p <.001 | p <.005 | p <.05 | p >.05

Table 1. Summary table showing a subset of results from the quantitative analysis in Experiments 1 and 2. Dark blue denotes $p < .001$, light blue denotes $p < .005$, and gray denotes $p < .05$. β represents the degree of change in trust or trend predictions between the compared conditions and t is β divided by its standard error.

4.2 Experiment 2

The goal of Experiment 2 was to understand the impact of including worst-case and best-case forecasts on viewers' trust. We hypothesized that including best- and worst-case forecasts that are visually distinct from the other forecasts (referred to in aggregate as the base forecast) would lower readers' trust (**H3A**), especially for visualizations that show fewer forecasts (**H3B**). We also hypothesized that after forecasts showed a sufficient number of models, participants would acknowledge the reliability of the base forecast and interpret extreme forecasts as useful for thoroughness instead of a threat to trustworthiness (**H3B**).

4.2.1 Quantitative Trust Results

We compared multiple linear-mixed effects equations to determine that the most parsimonious model specifies: an interaction between forecast visualizations (1-15 forecasts and CI 95) and best/worst-case forecasts (base forecasts, base forecasts + the best-case, and base forecasts + the worst-case), their lower order terms as fixed effects to predict variance in trust ratings, and random intercepts for each participant ($\chi^2(30) = 163.8$, $p = 0.00$). The R notation for this model is: $Trust \sim ForecastVis * BaseBestWorst + ForecastVis + BaseBestWorst + (1|Participant)$. We specified the 95% CI and base forecast conditions as the referents.

Omnibus Analysis. This analysis revealed significant interactions between *ForecastVis* * *BaseBestWorst* for 3 of 30 levels, suggesting that the relationship between the CI95 and the other forecast visualizations is different for the base vs worst-case and base vs best-case forecasts. To examine the nature of these interactions, we ran the same analysis on each condition separately. For each analysis, we focused on identifying if the 95% CI and the median stimuli were among the most trusted and identifying the asymptotic nature of trust (replicating Experiment 1). Due to the large number of significant effects, below we describe summaries of the analyses. Complete model outputs can be viewed in the supplemental materials.

Base Forecast Analysis. For the Base Forecast group, the analysis revealed that the 95% CI forecast was significantly more trusted than all other stimuli, as shown at the top of Figure 5 and Table 1 row F. The median ensemble line chart was the second most trusted, and was significantly more trusted than MFVs of 2, 4, 11, 12, 14, and 15 forecasts. However, we did not find clear evidence that trust approached an asymptote as the number of forecasts increased (i.e., no replication of **H1**), which may be due to the more narrow spread of forecasts, making the value of additional forecasts appear less consequential. As seen in Figure 5, the trust ratings for the MFVs indicate that seven forecasts result in the highest trust ratings. Table 2 demonstrates that participants could not accurately discriminate the number of forecasts visualized within this spread of forecasts, providing support for a shifted trade-off between thoroughness and clarity compared to Experiment 1 (see Figure 1 for visual comparisons).

Best- & Worst-Case Forecast Analyses. However, when the chart included best- or worst-case forecasts, trustworthiness increased with the number of forecasts up to a point (see middle and bottom of Figure 5;

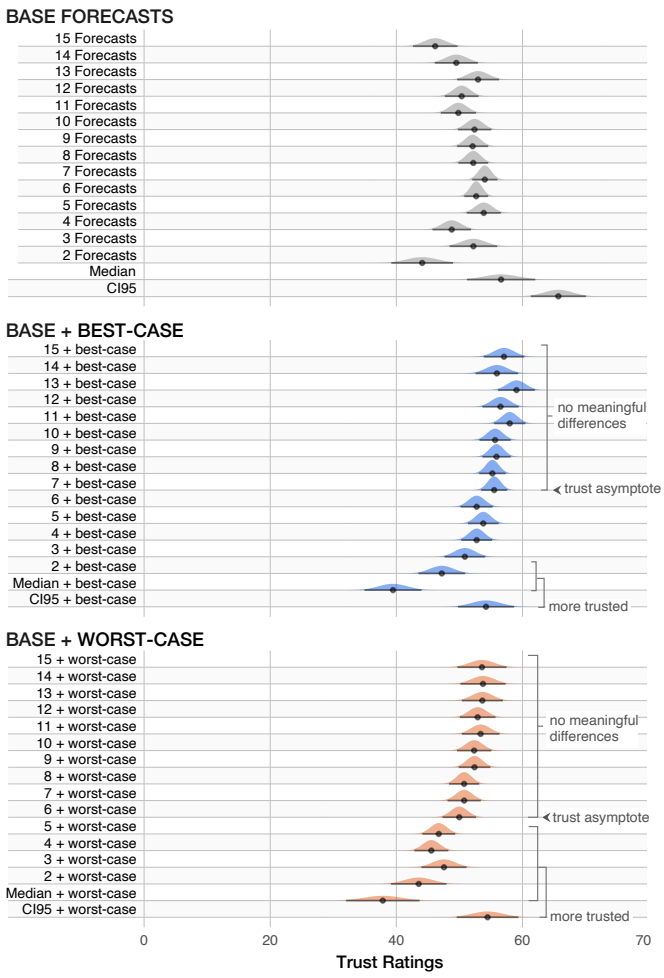

Fig. 5. Experiment 2 trust results. Annotations were added for less obvious significant effects.

H1 replicated). In accordance with Experiment 1, participants trusted the stimuli with best- or worst-case forecasts least when shown only the extreme forecast and one other forecast (a total of two forecasts).

Additionally, we found differences in the impact of adding best-case and worst-case forecasts to 95% CI visualizations. We found that when the best-case forecast was added to stimuli, the 95% CI stimulus was only more trusted than two of the other stimuli (annotated in Figure 5 middle and in Table 1 row G). On the other hand, when the worst-case forecast was added to a stimulus, the 95% CI was more trusted than six of the other stimuli (annotated in Figure 5 bottom and Table 1 row H).

Asymptote Analysis. Finally, we evaluated the location of the asymptotes that trust approaches using the same procedure described

in Experiment 1 (Section 4.1.1, Asymptote Analysis). For stimuli with additional best-case forecasts, trust plateaued after seven forecasts. For stimuli with worst-case forecasts, this plateau occurred after five forecasts were shown.

4.2.2 COVID-19 Trend Prediction

We conducted the same quantitative analysis as in Experiment 1 (Section 4.1.2) to understand the relationship between participants' interpretation of the COVID-19 mortality trend in the next two weeks and the stimuli (results shown in Figure 6). The results revealed a main effect where 95% CI was less likely to produce predictions that corresponded to the actual COVID-19 trend than all other visualization techniques except for the median (Table 1 row I; replicating Experiment 1). When changing the referent to the median forecast, we also found that the median was less likely to produce predictions that match the actual COVID-19 trend than the charts with more than two forecasts (Table 1 row J). Across the three conditions (base, base+worst, base+best), we did not find a visualization that consistently led to the most correct predictions.

We observed a main effect where the base forecasts (*Mean* = -.22) produced predictions more consistent with the subsequent COVID-19 trend than including worst-case forecasts (*Mean* = -.05, $\beta = .19$, $t(295) = 3.26$, $p = .001$), $CI[.08, .30]$). Further, graphs that included the best-case forecast (*Mean* = -.24) produced predictions more consistent with the actual COVID-19 trend than including the worst-case forecast ($\beta = .18$, $t(291) = 3.07$, $p = .002$, $CI[.06, .29]$), which makes sense, as COVID-19 mortalities dropped in the two weeks following the predictions and the best-case forecast indicated a decline in mortalities.

In line with Experiment 1, the base condition that showed two forecasts evoked the highest number of unsure responses (49 of 99), and the point-based forecast (6 of 99) and 95% CI forecast (5 of 99) had the least numbers of unsure responses. Unlike the base condition, both the stimuli with 95% CI forecasts and additional best-case or worst-case forecasts also elicited many unsure responses (best-case = 24 of 100, worst-case = 20 of 100). It appears that including the best or worst-case forecasts with confidence intervals decreases trust and makes people unsure of how to interpret the forecasts.

4.2.3 Manipulation Check

At the end of Experiment 2, participants counted the number of forecasts in eight stimuli (those that showed 8-15 forecasts). We included two more stimuli in the discrimination task for Experiment 2 because the forecasts covered a narrower range, resulting in more overplotting. We hypothesized that discrimination accuracy for stimuli showing eight or fewer forecasts would be nearly 100%. As shown in Table 2, we found that accuracy substantially decreased for 12-15 Forecasts, suggesting that for 12-15 Forecasts participants could not determine how many forecasts were shown, and that they were highly accurate in counting the eight forecasts.

4.3 Open Responses for Experiments 1 and 2

After reporting trust ratings, participants answered open-ended questions about their strategies in the trust rating task, why they thought there was uncertainty in COVID-19 forecasts, and why they believed there was disagreement between the models. To analyze these data, we had two independent raters read all 1500 responses and code them based on patterns in the responses. In the following sections, we report the three most frequent response-types/strategies for each question. A full breakdown of all the response strategies is available in supplemental materials. We computed inter-rater reliability scores for the three most frequent strategies. An inter-rater reliability score provides a metric for the level of agreement between raters. The inter-rater reliability (Cohen's Kappa) [17] was 91.2%, 89.1%, and 91.3% for questions 1, 2, and 3, respectively. Codes were not mutually exclusive, which will be apparent in the following passages that include example responses.

Question 1. "Please describe in as much detail as you can how you made your trust judgments." The three most common strategies used in the trust rating task, were 1) modifying trust ratings based on the number of forecasts shown, 2) rating trust based on the apparent

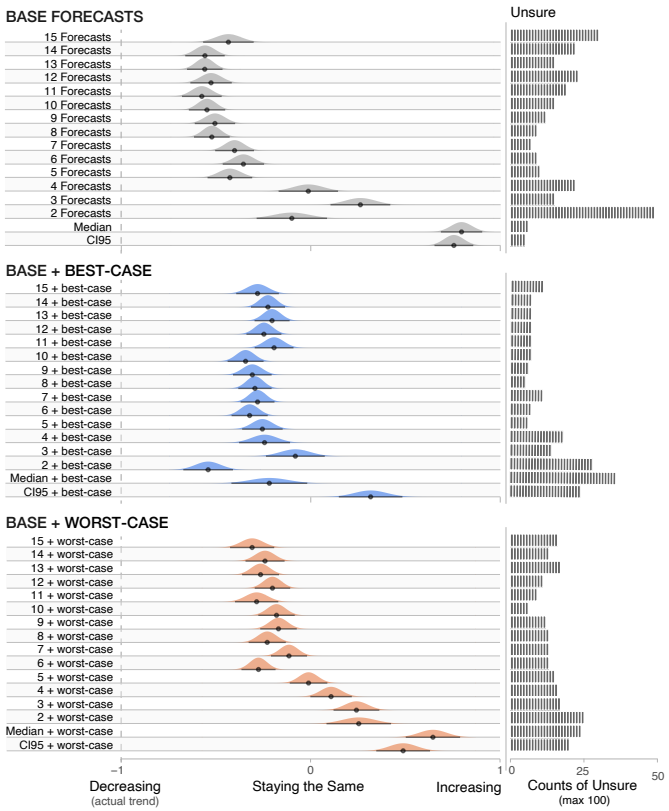

Fig. 6. Trend analysis for Experiment 2 where -1 is decreasing (actual trend). Right of the figure shows counts of "unsure" responses.

agreement or disagreement of the forecasts, and 3) using background knowledge about COVID-19 to determine the trustworthiness of the forecasts. Table 3 shows the proportions of the participants who used these strategies in each experiment.

Participants indicated that they calibrated their trust ratings to the number of forecasts (39.27%). Although this is the most common strategy reported, people incorporated the number of forecasts into their trust ratings in two distinct ways. The larger group of people indicated that they had greater trust for forecasts with more models (14.03%), for example, *"I perceived the graphs that included more models as more trustful."*

In contrast, a smaller proportion of people expressed the opposite strategy, where they rated forecasts with more models as less trustworthy (6.41%). For example, *"If there were many lines, it did not seem reliable to me because they did not give me concise information."* Participants who used this strategy often reported that increasing numbers of models made the visualizations confusing, such as, *"it is a little bit confusing if there are a lot of prediction lines. i think less lines, more trustworthy."*

The second most common strategy was to rate trust based on the perceived agreement or disagreement between the models. Participants mentioned focusing on either the model agreement, as in, *"I determined that a graph was trustworthy by counting the number of forecast models presented on each graph and checking to see if there was relative agreement among the various models regarding their forecasts"* or disagreement *"for me it is more trustworthy the ones that had less lines in the graph, when having too many lines in the graph the data presented seems chaotic and gives the image that the researchers doing the predictions are all on disagreement with each other."*

The third most common strategy was to rely on background knowledge to evaluate trust. Participants who reported using this strategy would primarily evaluate if the predictions matched their beliefs about the COVID-19 trend and rate those that confirmed their beliefs as more trustworthy. An example of this confirmation bias is, *"My trust judgements were made based on the fact that I expect Covid-19 numbers and*

Questions	Q1. Trust Judgement Strategy			Q2. Why Uncertainty in COVID-19 Forecasts			Q3. Why Differences Between Forecasts		
Response Codes	sensitivity to # of forecasts	focusing forecast agreement or disagreement	using background knowledge	human actions	stochastic processes	biological factors	different models or assumption	different data	different variables
Exp 1 Grayscale	36.5%	16.5%	18.5%	42.0%	40.0%	11.0%	32.0%	18.0%	17.5%
Exp 1 Color	42.5%	22.5%	13.5%	28.5%	40.0%	17.5%	29.5%	24.0%	17.5%
Exp 2 Base	32.3%	19.7%	24.2%	37.4%	27.8%	18.7%	25.8%	32.8%	15.2%
Exp 2 Worst-case	46.0%	18.0%	11.5%	33.5%	32.5%	14.5%	39.0%	28.5%	23.5%
Exp 2 Best-case	39.0%	26.5%	15.5%	34.5%	32.0%	24.5%	43.0%	21.5%	29.5%
Totals	39.3%	20.6%	16.6%	35.2%	34.5%	17.2%	33.9%	24.9%	20.6%

Table 3. Percent of participants in Experiments 1 and 2 who reported using each strategy or indicated a particular belief about COVID-19 forecasting. The table summarizes the three most reported codes per question and the percentages represent the averages of two raters.

deaths to increase over the next two weeks due to the start of the festive season, the opening of borders and the easing up on travel and other covid related restrictions. Thus, any forecast showing a prediction of Covid-19 deaths remaining the same or decreasing is, in my opinion, not trustworthy."

Some participants also identified conflicting strategies whereby they trusted both the simplistic graphs and those with more forecasts. In our trust ratings, we find evidence of these strategies in both high trust ratings for the median forecast and increasing trust with more forecasts. One participant explains, *"A graphic looks trustworthy to me if it either shows a great number of different models or only shows a single model. Different models present various possibilities and require more research. On the other hand, when a single model is presented, it somehow gives off the impression that there is more certainty regarding the prediction."* Also in line with the trust ratings, seven people reported that they found the 95% CI most trustworthy. For example, *"I did not trust graphs that were difficult to read (overcrowded lines). I liked the graph that had the confidence range displayed."*

Question 2. "Why do you think there is uncertainty in COVID-19 forecasts?" Before this study, readers' perceptions of the causes for uncertainty in COVID-19 forecasts were unclear. This information may be extremely useful for risk communicators who need to communicate unsure information.

The three sources of uncertainty in COVID-19 forecasts that participants cited the most were unpredictable human actions (e.g., not getting vaccinated or wearing masks, 35.1%), biological factors (e.g., new variants or mutations, 17.2%), and stochastic processes (e.g., general variability in forecasts, 34.4%). Uncertainty due to human actions and biological factors are contributors to uncertainty in COVID-19. In fact, some forecast models make different assumptions about mask and vaccine mandates, which contributes to the range of forecasts.

We used the code *stochastic process* when people described the difficulty of the forecasting process. These responses were often ill-defined, suggesting that the participants did not understand why there was uncertainty in the COVID-19 forecast but knew that it was there. For example a participant wrote, *"Because it is impossible to predict the future and the virus is even more unpredictable."*

Question 3. "Why do you think the various forecasts made different predictions about the COVID-19 deaths in the US?" Participants identified the most common causes of disagreement between forecasts as different models with various assumptions (33.8%), different datasets (24.9%), and different variables (20.6%). As with the prior codes, these were not mutually exclusive, and people commonly cited each of these causes as contributors to forecast disagreement. For example, participants wrote, *"They may have used different variables, sample sizes, data, methodologies"* or *"Different models have different sources of data, variables, and weights to those variables. These various differences will create a natural variety in models..."* Such responses suggest a relatively sophisticated understanding of forecast variability.

However, not all participants had a deep understanding of the factors that contribute to forecast disagreement. Indeed, one of the most commonly reported assumptions was that each forecast used data from different states (35 people), even though the label on the graph indicated that the forecasts were for the whole United States. One participant wrote, *"I honestly don't have a clue and I feel really stupid right now for that very reason. It probably has something to do with different data taken into account or maybe making predictions for different states. I'm guessing here."* Twenty-two people stated outright that they did not know why the forecasts made different predictions. In contrast, so few people indicated that they were unsure for the other two questions that we did not include a code for unsure.

4.3.1 Discussion

Experiments 1 and 2 found that trust generally increases when charts display more models (**H1A**) and plateaus around 6-9 forecasts (**H1B**). We also found that the likelihood of predicting that the COVID-19 mortalities would trend in a direction consistent with the actual trend increases with the number of forecasts and plateaus around 5-7. Accuracy in counting the number of forecasts depends on the spread of the models, and participants could accurately discern 8-10 lines.

Consistently, people had high trust in visualizations that showed less visual information, including 95% CI, a single median forecast, and grayscale encoded forecasts. The 95% CI and the median forecast also consistently produced poor performance in the trend judgment task.

Participants corroborated the empirical findings by describing their greater trust for forecasts with numerous models while also trusting the more simplistic visualizations. Further, they had a robust understanding of uncertainty sources in COVID-19 predictions but had some incorrect assumptions about why scientists produce forecasts that disagree.

4.4 Experiment 3

Given the consistently high trust evoked by the 95% CI, we wanted to test if part of that trust was due to the label indicating "95%." To test if participants reasonably scale their trust with the reported range of the confidence interval, we conducted a follow-up experiment in February 2022. We asked four groups of participants to rate the trustworthiness of the exact same COVID-19 forecast visualization from the week of February 5th, 2022 (Figure 2). However, we changed the figure caption to read as 25, 50, 95, or 99% confidence interval. We also tested a second manipulation where we added direct annotations with these same values to the forecasts (see Figure 2). This experiment was a follow-up study, which we did not preregister. Therefore, the analysis was an exploratory post hoc investigation of the trends. The linear equation we used to model this experiment included the interaction between reported interval size (25, 50, 95, and 99%) and labeling (caption vs caption and annotation) and their lower order terms to predict variance in trust ratings. The R notation for this model is: $Trust \sim IntervalSize * Labeling + IntervalSize + Labeling$. We specified the CI25 and the caption-only conditions as the referents.

As shown in Figure 7, we found no meaningful impact of adding the direct annotations (β = -0.41, $t(792)$ = -0.13, p = 0.90, CI[-6.6, 5.8]). There was also no interaction between reported size of the interval and the labeling. However, we did find a main effect of the reported interval size such that both 25% and 50% were less trusted than 99% (25CI: β = 15.27, $t(792)$ = 4.84, p = .00, CI[9.08, 21.5]) (50CI: β = 11.61, $t(792)$ = 3.68, p = .0002, CI[5.42, 17.80]). Further, 25% and 50% were also less trusted than 95% (25CI: β = 16.89, $t(792)$ = 5.36, p = .00, CI[10.70, 23.08]) (50CI: β = 13.23, $t(792)$ = 4.19, p = 000, CI[7.04, 19.42]). However, there were no differences between 25% and 50% (β = 3.66, $t(792)$ = 1.16, p = 0.25, CI[-2.53, 9.85]) or 95% and 99% (β = -1.62, $t(792)$ = -0.51, p = 0.61, CI[-7.81, 4.57]).

Fig. 7. Experiment 3 trust results, for which the stimuli with captions are encoded with orange and those with additional annotations are in blue.

4.4.1 Discussion

This analysis revealed that the COVID-19 forecast labeled as 25% confidence (*Mean* = 50.67) and 50% confidence (*Mean* = 52.9) were both rated at an average of roughly 50 on the trust scale. Rating the 50% CI as a 50 on the scale from 1-100 is reasonable, but it is unclear why participants did not rate the graph with the 25% CI caption lower than the one indicating 50% CI. One explanation is that all of the intervals were physically the same size. However, if this hypothesis were true, those who viewed the 95% (*Mean* = 70.3) and 99% (*Mean* = 66.5) should not have increased trust ratings. Further, it is unclear why the CI95 and CI99 were not rated higher than 70 on the trust scale.

Viewers are sensitive to the label indicating interval size, but they do not respond to this information reasonably or consistently. It is particularly concerning that participants placed high trust in the confidence interval labeled as 25%. Keeping Experiments 1 and 2 in mind, the trust ratings for the confidence intervals labeled as 25% and 50% are on par with the most trusted MFVs we tested.

5 General Discussion

This work indicates that visualization design choices easily influence trust, and trust is not indicative of successful interpretation. While we all may aim to "trust the trustworthy, but not the untrustworthy", as Philosopher Onora O'Neill writes, poorly designed forecast visualizations can confound our ability to successfully calibrate trust. For example, visualizations that show no uncertainty evoke unduly high trust while producing poor decision quality. Despite their complicated relationship in forecast visualizations, our findings point to several methods to balance performance and trust. These findings include:

1. Increasing forecasts in MFVs generally increase trust until a plateau around 6-9 forecasts (Sections 4.1.1 and 4.2.1), which also results in trend predictions most consistent with the future event (Sections 4.1.2 and 4.2.2). The specific location of a trust/performance trade-off will likely vary by data, visual encoding techniques, and target audience. Our findings suggest that trust-performance trade-off points will near this maximum discernible forecast number, but designers should consider nuances in their applications when selecting number of forecasts.
2. Single point-based forecasts and 95% CIs evoke trend predictions that are *least* consistent with the data (Sections 4.1.2 & 4.2.2), whereas their trust is the highest (Sections 4.1.1 and 4.2.1).
3. People report increasing their trust with the number of models shown but also report perceived distrust when shown too many models at once (Section 4.3).
4. Viewers do not appropriately scale their trust to the indicated confidence interval size (Section 4.4).

Participants trusted visualizations they viewed as clearer, which often showed less complex visual information, like the point-based forecast and 95% CI. This finding corresponds with prior work indicating that clarity is a key component of trust [59] and that viewers do not require extensive uncertainty information to trust a visualization [23, 38]. Although people may trust visualizations that appear to provide clear information by displaying less information, lacking knowledge of the full range of outcomes can ultimately harm their judgments. Further, participants' trust in confidence intervals did not scale proportionally to labelled interval size, suggesting that viewers are unclear how to interpret confidence intervals despite highly trusting them. A combination of high trust and misinterpretation can be dangerous and adds to the mounting evidence that viewers have difficulty interpreting confidence intervals (e.g,. [8, 10, 19, 21, 31, 57]).

We also found that participants had difficulty interpreting MFVs with best- and worst-case forecasts. A high number of participants were unsure how to interpret 95% CIs when paired with extreme forecasts, and extreme forecasts often incorrectly biased MFV trend predictions.

In examining open-ended responses, we found that participants have varied ideas about the definition of trust, what makes something trustworthy, and the attributes that contribute to trust. These discrepancies highlight the challenge of examining trust, especially in an ecologically valid context (e.g., closely matching the conditions of an experiment to real-world situations [48]). Although we offer a less controlled treatment of trust, this work is unique because the findings offer critical insights into how the US public thinks about trust in information visualizations during the real global health crisis impacting them presently.

5.1 Limitations and Caveats

Work on visualization interpretation is commonly conducted in a highly controlled fashion with fictional events, fabricated data, or situations that do not directly affect the viewers. Such work reveals meaningful insights about readers' perception but has limited ecological validity. We chose to sacrifice some experimental control to prioritize ecological validity. Thus, there are various limitations and caveats to our approach.

To support clarity in our statistical reporting, we chose to evaluate participants' predictions of future COVID-19 trends based on their consistency with actual COVID-19 mortalities two weeks after the forecast. However, evaluating the accuracy of participant predictions may be an oversimplification. We concluded that the actual 32.4% reduction in mortalities provided cause for "decreasing" to be considered the only correct participant prediction. However, a certain level of decreasing mortalities could allow "staying the same" to also be acceptable.

We also presented several statistical results that were not fully consistent. In particular, we chose to show stimuli that ranged from 1-15 forecasts, and, at times, forecasts would produce results that were not consistent with the trend. The variability could be due to many causes, as we did not generate the data or models. To examine the consistency of our findings, future work is needed to experimentally control data and manipulate the properties of forecasts to determine sources of this variability. Nevertheless, general consistency in trends across our experiments and consistency with prior research provides evidence that our findings are reliable.

Finally, we did not examine many other factors that may influence trust in COVID-19 forecasts. For example, confidence in visualizations' authors [65], confirmation bias, and individual differences likely play an important role in visualization trust. Similarly, trust may be further impacted by untested design decisions. We focused on three commonly manipulated design variables: color, number, and shape of forecasts shown. The impact of other design choices, such as line thickness, presence of markers, line opacity, and even the context in which visualizations are shown (e.g., on paper, via AR integration), have the potential to impact audience trust. We need future work to systematically evaluate the multiple facets of trust and distrust [20, 35, 42, 44, 59] in information visualizations, and forecast visualizations specifically. Further, this work did not define trust for participants or have converging measures of trust, which adds ambiguity to the results.

6 Conclusions

Visualizations can aid the public's understanding of and decision-making during pandemics, or they can confuse and seed distrust. Here, we present methods for improving trust while maintaining trend interpretation performance via multiple forecast visualization (MFV) guidelines. In addition to communicating important health information, conveying uncertainty in public-facing visualizations increases general understanding and tolerance of variability in future events.

Acknowledgments

The National Science Foundation (# 2028374) support this work.

REFERENCES

[1] J. Agley, Y. Xiao, E. Thompson, and L. Golzarri-Arroyo. Using infographics to improve trust in science: a randomized pilot test. *BMC Research Notes*, 14, May 2021. doi: 10.1186/s13104-021-05626-4

[2] Y. Algan and P. Cahuc. Trust, growth, and well-being: New evidence and policy implications. In *Handbook of Economic Growth*, vol. 2, pp. 49–120. Elsevier, 2014.

[3] O. Bargain and U. Aminjonov. Trust and compliance to public health policies in times of COVID-19. *Journal of Public Economics*, 192:104316, 2020.

[4] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01

[5] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005.

[6] A. P. Boone, P. Gunalp, and M. Hegarty. Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of experimental psychology: applied*, 24(3):275, 2018.

[7] W. C. Casper, B. D. Edwards, J. C. Wallace, R. S. Landis, D. A. Fife, et al. Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology*, 105(4):390, 2020.

[8] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla. Examining effort in 1D uncertainty communication using individual differences in working memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):411–421, 2021.

[9] F. L. Cook, L. R. Jacobs, and D. Kim. Trusting what you know: Information, knowledge, and confidence in social security. *The Journal of Politics*, 72(2):397–412, 2010.

[10] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014.

[11] E. Y. Cramer, Y. Huang, Y. Wang, E. L. Ray, M. Cornell, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, K. House, D. Jayawardena, A. H. Kanji, A. Khandelwal, K. Le, J. Niemi, A. Stark, A. Shah, N. Wattanachit, M. W. Zorn, N. G. Reich, and U. S. C. F. H. Consortium. The United States COVID-19 Forecast Hub dataset. *medRxiv*, 2021. doi: 10.1101/2021.11.04.21265886

[12] S. Deitrick and E. A. Wentz. Developing implicit uncertainty visualization methods motivated by theories in decision science. *Annals of the Association of American Geographers*, 105(3):531–551, 2015.

[13] S. Dryhurst, C. R. Schneider, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, D. Spiegelhalter, and S. Van Der Linden. Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, 23(7-8):994–1006, 2020.

[14] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.

[15] J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, Third ed., 2019.

[16] F. Fukuyama. *Trust: The social virtues and the creation of prosperity*. Simon and Schuster, 1996.

[17] M. Gamer, J. Lemon, I. Fellows, and P. Singh. *irr: Various Coefficients of Interrater Reliability and Agreement*, 2019. R package version 0.84.1.

[18] R. G. Gelos and S. J. Wei. Transparency and international investor behavior. Technical report, National Bureau of Economic Research, 2002.

[19] M. Grounds, S. Joslyn, and K. Otsuka. Probabilistic interval forecasts: An individual differences approach to understanding forecast communication. *Advances in Meteorology*, 2017:1–18, 09 2017. doi: 10.1155/2017/3932565

[20] R. S. Gutzwiller, E. K. Chiou, S. D. Craig, C. M. Lewis, G. J. Lematta, and C.-P. Hsiung. Positive bias in the 'Trust in Automated Systems Survey'? An examination of the Jian et al. (2000) scale. In *Proceedings of the Human Factors and Ergonomics Society annual meeting*, vol. 63, pp. 217–221. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[21] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020. doi: 10.1145/3313831.3376454

[22] C. Hood and D. Heald. *Transparency: The key to better governance?*, vol. 135. Oxford University Press for The British Academy, 2006.

[23] J. Hullman. Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, 2019.

[24] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):446–456, 2017.

[25] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS One*, 10(11), 2015.

[26] S. Joslyn and J. LeClerc. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013.

[27] S. Joslyn, L. Nemec, and S. Savelli. The benefits and challenges of predictive interval forecasts and verification graphics for end users. *Weather, Climate, and Society*, 5(2):133–147, 2013.

[28] S. Joslyn and S. Savelli. Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Frontiers in Computer Science*, 2:58, 2021. doi: 10.3389/fcomp.2020.590232

[29] M. F. Jung, D. Sirkin, T. M. Gür, and M. Steinert. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2201–2210. ACM, 2015.

[30] J. Kahn. Reporting statistics in APA style. *American Psychological Association*, 2010.

[31] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 10 2020. doi: 10.1109/TVCG.2020.3030335

[32] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2018.

[33] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103, 2016.

[34] M. Kay, D. Morris, M. Schraefel, and J. A. Kientz. There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 401–410, 2013.

[35] S. Y. Komiak and I. Benbasat. A two-process view of trust and distrust building in recommendation agents: A process-tracing study. *Journal of the Association for Information Systems*, 9(12):2, 2008.

[36] C. A. Latkin, L. Dayton, G. Yi, A. Konstantopoulos, and B. Boodram. Trust in a COVID-19 vaccine in the US: A social-ecological perspective. *Social Science & Medicine (1982)*, 270:113684, 2021.

[37] C. Lee, T. Yang, G. D. Inchoco, G. M. Jones, and A. Satyanarayan. Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox science online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2021.

[38] D. Leffrang and O. Müller. Should I follow this model? The effect of uncertainty visualization on the acceptance of time series forecasts. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 20–26, 2021. doi: 10.1109/TREX53765.2021.00009

[39] S. Leo. Mistakes, we've drawn a few: Learning from our errors in data visualisation. `https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368`, 2019. Online; accessed 24-June-2022.

[40] L. Liu, A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2165–2178, 2016.

[41] L. Liu, L. Padilla, S. H. Creem-Regehr, and D. H. House. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):882–891, 2018.

[42] S. Marsh and M. R. Dibben. Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In *International Conference on Trust Management*, pp. 17–33. Springer, 2005.

[43] S. Maykrantz, T. Gong, A. Petrolino, B. Nobiling, and J. Houghton. How trust in information sources influences preventative measures compliance during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 18:5867, 05 2021. doi: 10.3390/ijerph18115867

[44] E. Mayr, N. Hynek, S. Salisu, and F. Windhager. Trust in information visualization. In *TrustVis@ EuroVis*, pp. 25–29, 2019.
[45] A. Oldeweme, J. Märtins, D. Westmattelmann, and G. Schewe. The role of transparency, trust, and social influence on uncertainty reduction in times of pandemics: Empirical study on the adoption of COVID-19 tracing apps. *J Med Internet Res*, 23(2):e25893, Feb 2021. doi: 10.2196/25893
[46] L. Padilla, H. Hosseinpour, R. Fygenson, J. Howell, R. Chunara, and E. Bertini. Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Scientific reports*, 12(1):1–14, 2022.
[47] L. Padilla, M. Kay, and J. Hullman. *Uncertainty Visualization*, pp. 1–18. Wiley StatsRef: Statistics Reference Online, 2021. doi: 10.1002/9781118445112.stat08296
[48] L. M. Padilla. A case for cognitive models in visualization research: Position paper. In *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 69–77. IEEE, 2018.
[49] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1):1–16, 2017.
[50] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *IEEE Symposium on Information Visualization*, pp. 89–96. IEEE, 2004.
[51] Plotly Technologies Inc. Collaborative data science, 2015.
[52] *Prolific*. (2014), Prolific, Ltd. [Online]. Available: `https://www.prolific.co`.
[53] *Fraud detection*. Qualtrics, LLC (2014). Qualtrics [software]. Utah, USA: Qualtrics.
[54] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
[55] N. Reich. The COVID-19 forecast hub, 2021.
[56] N. Reinholtz, S. J. Maglio, and S. A. Spiller. Stocks, flows, and risk response to pandemic data. *Journal of Experimental Psychology: Applied*, 27(4):657, 2021.
[57] L. F. Rinne and M. M. Mazzocco. Inferring uncertainty from interval estimates: Effects of alpha level and numeracy. *Judgment and Decision Making*, 8(3):330, 2013.
[58] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016.
[59] A. K. Schnackenberg and E. C. Tomlinson. Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management*, 42(7):1784–1810, 2016.
[60] C. R. Schneider, A. L. Freeman, D. Spiegelhalter, and S. van der Linden. The effects of quality of evidence communication on perception of public health information about COVID-19: Two randomised controlled trials. *PLOS One*, 16(11):e0259048, 2021.
[61] C. G. Sibley, L. M. Greaves, N. Satherley, M. S. Wilson, N. C. Overall, C. H. Lee, P. Milojev, J. Bulbulia, D. Osborne, T. L. Milfont, et al. Effects of the COVID-19 pandemic and nationwide lockdown on trust, attitudes toward government, and well-being. *American Psychologist*, 75(5):618, 2020.
[62] A. M. Van Der Bles, S. van der Linden, A. L. Freeman, and D. J. Spiegelhalter. The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, 117(14):7672–7683, 2020.
[63] R. J. D. Vergara, P. J. D. Sarmiento, and J. D. N. Lagman. Building public trust: A response to COVID-19 vaccine hesitancy predicament. *Journal of Public Health*, 43(2):e291–e292, 2021.
[64] C. Xiong, L. Padilla, K. Grayson, and S. Franconeri. Examining the components of trust in map-based visualizations. In R. Kosara, K. Lawonn, L. Linsen, and N. Smit, eds., *EuroVis Workshop on Trustworthy Visualization (TrustVis)*. The Eurographics Association, 2019. doi: 10.2312/trvis.20191186
[65] Y. Zhang, N. Suhaimi, N. Yongsatianchot, J. D. Gaggiano, M. Kim, S. A. Patel, Y. Sun, S. Marsella, J. Griffin, and A. G. Parker. Shifting trust: Examining how trust and distrust emerge, transform, and collapse in COVID-19 information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501889
[66] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker. Mapping the landscape of COVID-19 crisis visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445381