AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

© 2025 The Author(s)
ISSN: 2325-9965

2025, Vol. 12, No. 3, 268–285
https://doi.org/10.1037/dec0000260

Visual Displays for Communicating Multiple Uncertain Forecasts

Yanran Yang[1, 2], Gabrielle Wong-Parodi[3, 4, 5], and Baruch Fischhoff[1, 6]

[1] Department of Engineering and Public Policy, Carnegie Mellon University
[2] Global Health Research Center, Duke Kunshan University
[3] Department of Earth System Science, Stanford University
[4] Stanford Woods Institute for the Environment, Stanford, California, United States
[5] Department of Environmental Social Sciences, Stanford University
[6] Institute for Politics and Strategy, Carnegie Mellon University

When decision makers receive multiple uncertain forecasts (as with the COVID-19 pandemic), they need displays that will help them to integrate the competing predictions. One display option is presenting the full *suite* of forecasts. A second is presenting a summary *ensemble*, reducing cognitive load at the price of obscuring disagreements. One compromise is presenting both, allowing decision makers to tailor their usage, with an even greater cognitive load. A second compromise is presenting one display but allow users to click through to see the other. In two experiments, using a suite of 10 CDC forecasts and an ensemble based on median values, we compared users' performance with displays presenting the (a) suite, then adding the ensemble; (b) the ensemble, then adding the suite; and (c) the suite and the ensemble combined. Using diverse, online convenience samples and hypothetic scenarios, we found that participants responded similarly to the combined displays, if they saw them initially together or only after seeing the suite or the ensemble. Participants responded differently if they saw just the suite or just the ensemble—as would happen if they did not click through from the single display. Outliers in the suite display appeared responsible for the difference. The result was robust whether the suite displays were ordered randomly, by confidence interval widths, or by median prediction. Comprehension was good with all displays. These results indicate the importance of ensuring that decision makers receive and understand both the suite and the ensemble perspective when there are multiple uncertain forecasts.

Keywords: forecasts, data visualization, risk communication, uncertainty, updating

Supplemental materials: https://doi.org/10.1037/dec0000260.supp

David V. Budescu served as action editor.

Yanran Yang https://orcid.org/0000-0002-1065-4863

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article. The authors thank Christine Cao at Department of Engineering & Public Policy, Carnegie Mellon University for her help with the execution of Study 2.

This article was supported by the Department of Engineering and Public Policy, Carnegie Mellon University and the Centers for Disease Control and Prevention (CDC) and the U.S. Department of Health and Human Services (HHS) as part of financial assistance award U01IP001121 totaling $3,412,979, with 100% funded by CDC/HHS. The contents are those of the author(s) and do not necessarily represent the official view of, nor an endorsement, by CDC/HHS, or the U.S. Government.

Yanran Yang played a lead role in data curation,

continued

With COVID-19, as with other important risks, machine learning has spurred multiple teams to estimate the same risks, using somewhat different methods, data, and assumptions. A practical question, when displaying those forecasts, is how to convey both the uncertainty of the individual risk estimates and the degree of agreement among them (Toet et al., 2019). We demonstrate a general method for evaluating alternative displays, illustrated with the multiple forecasts that the Centers for Disease Control and Prevention (CDC) provided to aid government, health, and general public decision makers. Each week during the height of the pandemic, CDC received short-term (1-,2-,3-, and 4-week) forecasts for county, state, and national COVID-19 deaths, hospital admissions, and cases from more than 20 modeling teams. CDC continues these practices with hospitalization forecasts for seasonal influenza and COVID-19. CDC displays the *suite* of individual forecasts, along with an *ensemble* reflecting the medians of the individual forecasts. Many methods have been tested for combining forecast distributions, including the simple average, the median, the envelope of the lowest low and highest high estimate, and various symmetric and asymmetric (interior and exterior) trimmed sets (Huang et al., 2024; Park & Budescu, 2015). Among these methods, the median has proven a robust aggregator (Gaba et al., 2017; Grushka-Cockayne & Jose, 2020). Indeed, CDC's median-based ensemble models have been better calibrated than most individual models for forecasting COVID-19 deaths, in terms of observations falling within prediction intervals (Cramer et al., 2022).

CDC relies on visual displays to communicate its forecasts, as have forecasters of hurricanes (Broad et al., 2007), weather (Stephens et al., 2019), and inflation (Aikman et al., 2011). Figure 1 shows one CDC visualization (from 2021 September). Like all displays, these reflect many design decisions, whose cumulative effect can only be known through empirical evaluation (Woloshin et al., 2023). After briefly reviewing research on communicating the uncertainty of individual forecasts, we evaluate the effects of several design decisions on conveying uncertainty within and agreement across forecasts.

Visual Display of Model Uncertainty

One common way of displaying the uncertainty of alternative hurricane forecasts is with a cone representing some fraction of the pooled forecasts. Boone et al. (2018) found that residents express greater intentions to evacuate if they live inside the cone of hurricane forecasts. Ruginski et al. (2016) contrasted lay evaluations of suites of individual hurricane paths and ensembles represented by uncertainty cones. They found that participants predicted greater, and later, hurricane damage, when the cones were wider, apparently inferring intensity from uncertainty. Participants shown suites of individual forecast paths expected less damage and saw less uncertainty. Padilla et al. (2017, 2022) similarly found that people interpreted suites of forecasts as limiting the range of possibility (and uncertainty). Witt and Clegg (2022) found that participants were less likely to evacuate when shown dynamic visualizations for suites of individual hurricane forecast paths. Demeritt et al. (2007) found that experts were more likely to "wait and see," before issuing evacuation orders, when shown suites of individual flooding forecasts.

Thus, we predict that suites of individual forecasts will convey greater uncertainty and impel more cautious decisions than will aggregated forecast ensembles. When displays induce different responses, forecasters must decide which to present. Greis et al. (2017) found that, with weather forecasts, participants preferred suites to ensembles and to receive both side-by-side, rather than receiving either display separately. That was especially true for decision scenarios with higher stakes (e.g., weddings and travel). However, the displays in that study were much simpler than uncertain forecasts like those in Figure 1, which have both intra- and intermodel uncertainty. Some people might be overwhelmed by one, much less two, such displays.

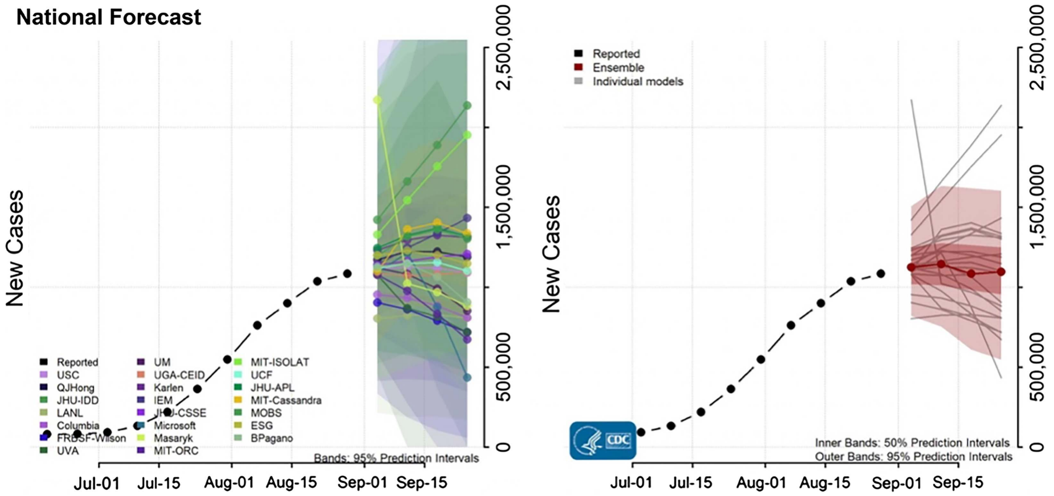
Here, we contrasted responses to a suite of individual forecasts and the associated ensemble,

formal analysis, visualization, and writing–original draft and an equal role in conceptualization, methodology, and writing–review and editing. Gabrielle Wong-Parodi played a supporting role in formal analysis and supervision and an equal role in conceptualization, methodology, and writing–review and editing. Baruch Fischhoff played a lead role in supervision and funding acquisition and an equal role in conceptualization, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Yanran Yang, Department of Engineering and Public Policy, Carnegie Mellon University, 5215 Wean Hall, PA 15213, United States. Email: yanrany@andrew.cmu.edu

Figure 1

Screenshot Example of CDC's Visualization for National COVID-19 Case Forecasts



Note. Left: Suite of forecasts from all eligible models, with the black line denoting actual case numbers and colored lines denoting each model's predicted cases for the next 4 weeks and colored bands denoting the 95% prediction intervals. Right: Ensemble model, with the black line denoting actual case numbers, the red line denoting the ensemble model's predicted cases for the next 4 weeks, and the red band denoting the 95% prediction interval. The gray lines denote the cases each model predicted, as do the colored lines in the left graph. CDC = Centers for Disease Control and Prevention. See the online article for the color version of this figure.

presented simultaneously or sequentially—as would happen if users landed on a page with one display and exercised the option to "click through" and examine the second (Wong-Parodi & Strauss, 2014). Studies of how people revise their judgments after receiving additional expert advice, in numeric form, typically find that initial information receives undue weight (Mannes et al., 2012; Yaniv, 2004a, 2004b; Yaniv & Kleinberger, 2000; Yaniv & Milyavsky, 2007). We expected a similar pattern with visual displays, which we do not believe have been studied previously.

In research on advice giving and taking, Harries et al. (2004) found that individuals tend to focus on the consensus opinion when aggregating if they are less experienced, but give greater weight to extreme opinions as they grow more experienced. Ocean (2024) found that when judging product quality with numeric ratings, merely flagging suspiciously extreme ratings barely affected consumers' judgments. As Pinker (1990) suggested, visual salience was an important factor in determining what information to extract from the graphs, as a bottom-up process. Within suites of COVID-19 forecasts, we predict that outliers will be visually salient, unlike numerical displays—where they are "just numbers."

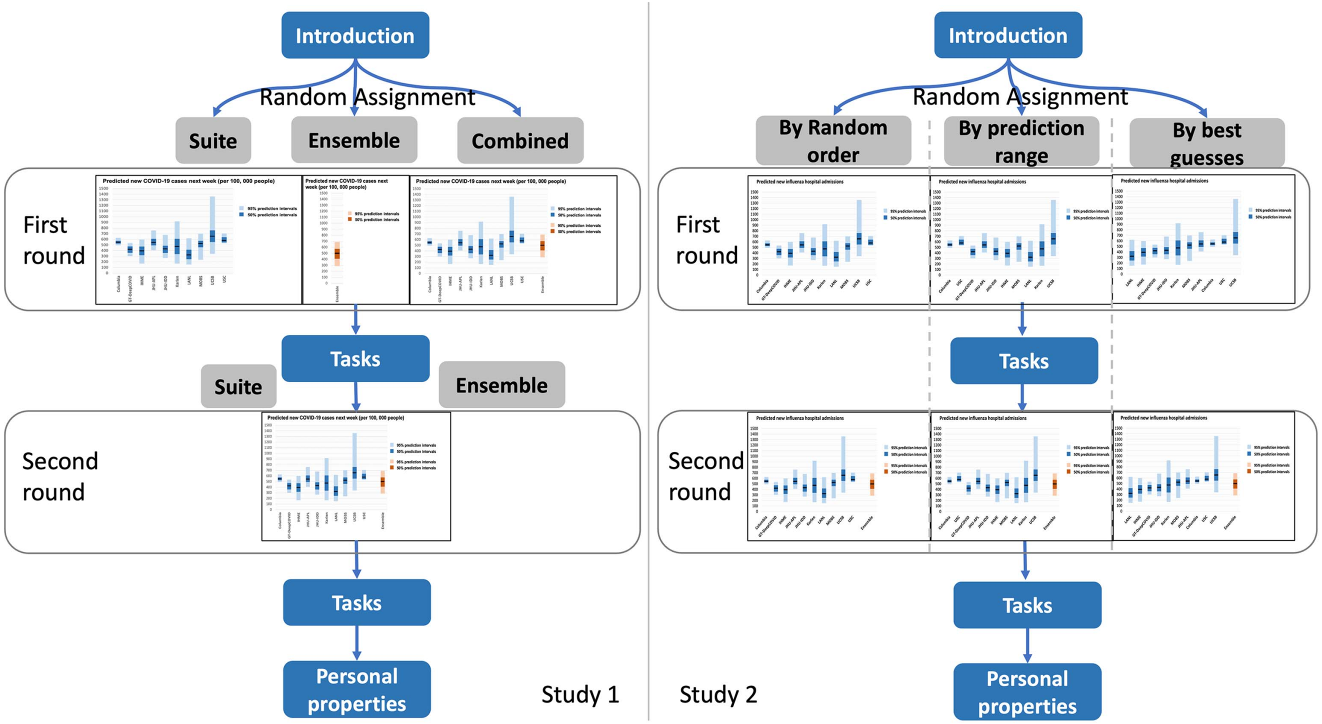
Design

Study 1 compares nonexpert judgments and decisions based on graphic displays of COVID-19 forecasts that express uncertainty either as a suite of 10 individual forecasts or as an ensemble summary of those forecasts, presented either sequentially or simultaneously (see Figure 2). Study 2 compares judgments and decisions based on displays presenting the same suite of 10 individual forecasts, either in a random order or ordered by either the range or the median value of the forecasts (see Figure 2 Right).

For our task environment, we created a hypothetical situation, asking participants to help friends decide whether to send their child back to school or be given the flu vaccine (Study 2). They are asked to (a) interpret the forecast visualizations for their friends and (b) suggest what they should do, based on the forecasts and their friends' risk tolerance. Participants then (c) indicate their confidence in their interpretation and decision advice and (d) rate

Figure 2

Schematic of the Experiment Design: Study 1 (Left) and Study 2 (Right)



Note. See the online article for the color version of this figure.

the display's helpfulness. This design provides a common risk tolerance for all participants, that of their (hypothetical) friends. Our stimuli are 10 actual forecasts produced by different CDC modeling teams, whose names are displayed but not explained (see Figure 3).

Study 1

We compare responses to the five ways in which users might encounter these displays: (a) *ensemble;* (b) *suite*; (c) *combined* (ensemble and suite); (d) *combined-ensemble first*, both displays after initially seeing the ensemble; and (e) *combined-suite first*, both displays after initially seeing the suite. We call any change between responses to the single and combined display *updating*.

Initial Judgment and Decisions

We expect outliers to drive differences between the displays, as with individual storm paths for hurricane forecasts. We define outliers as forecasts with unusually high or low values, within a suite. For the forecasts in our displays, the outliers are all unusually high values (Karlen, University of California, Santa Barbara in Figure 3), as the forecasts have similar lower bounds. Thus, we predict that participants initially shown the suite and combined displays will perceive greater uncertainty and suggest more cautious decisions. That uncertainty will emerge in three ways.

> *Hypothesis 1:* Suite and combined group participants will give higher initial estimates for the midpoint of the 95% confidence interval for predicted COVID-19 cases than will ensemble participants.

> *Hypothesis 2:* Suite and combined group participants will give wider initial confidence intervals for predicted COVID-19 cases than will ensemble group participants.

> *Hypothesis 3:* Suite and combined group participants will be less likely to suggest that their friends send their child back to school (risk avoidant) than will ensemble group participants.

Updating (Ensemble and Suite Group)

If, as predicted, the suite display creates greater uncertainty, adding the ensemble in the combined display should reduce it, unless participants are firmly anchored in their initial responses. The opposite should be true for participants who receive the ensemble first:

> *Hypothesis 4:* Ensemble group participants' final confidence intervals will be wider than their initial ones.

> *Hypothesis 5:* Ensemble group participants' final midpoint estimates will be higher than their initial ones.

> *Hypothesis 6:* Suite group participants' final confidence intervals will be narrower than their initial ones.

> *Hypothesis 7:* Suite group participants' final midpoint estimates will be lower than their initial ones.

We also hypothesized that adding a second display will increase the confidence of participants who initially see just one display.

> *Hypothesis 8:* Suite and ensemble group participants will have greater confidence in their judgments and decision advice when the second display is added and they will rate the full display as more helpful than the initial one.

Final Judgments and Decisions

Hullman et al. (2017) found no consistent pattern in whether participants prefer to start with more or less aggregated displays. As a result, we had no prediction regarding order effects (Poulton & Freeman, 1966).

> *Hypothesis 9:* Ensemble, suite, and combined group participants will respond similarly to the full display.

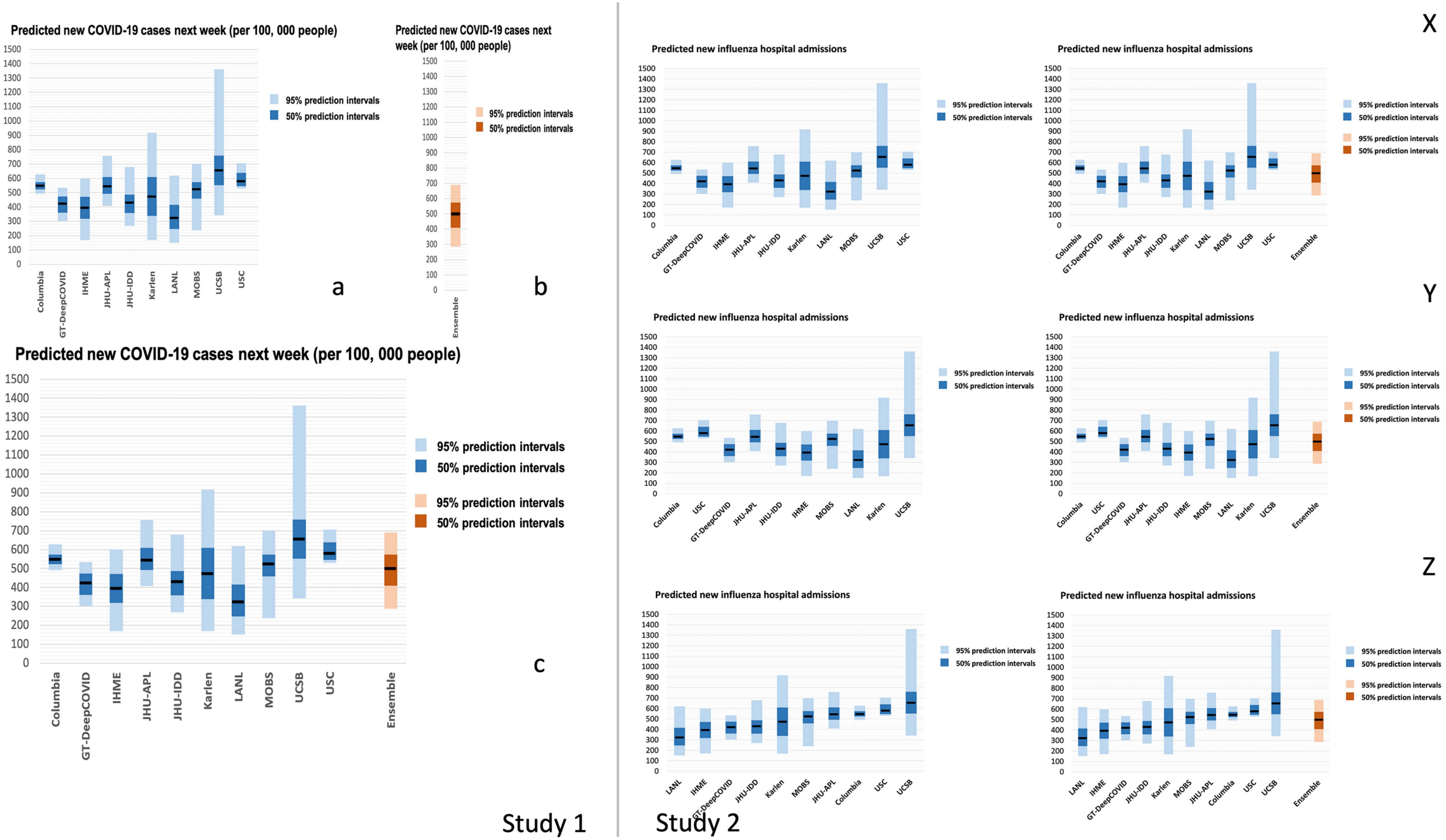
Display Evaluations

We predicted that participants would prefer the suite display's fuller picture and have greater confidence in judgments and decisions made with it:

> *Hypothesis 10:* Suite and combined group participants will be more confident in their estimates (risk judgments) and decision advice than will ensemble group participants.

Figure 3

Graphic Presentation of Uncertain Forecasts Used in Study 1 (Left) and Study 2 (Right)



Note. Study 1: (a) suite of 10 forecasts, (b) ensemble of the 10 forecasts, and (c) suite and ensemble combined. Study 2: (X) suite of 10 forecasts in random order followed by suite and ensemble combined, (Y) suite sorted by prediction ranges followed by combined, and (Z) suite sorted by best guesses followed by combined. See the online article for the color version of this figure.

> *Hypothesis 11:* Suite and combined group participants will rate their display as more helpful than will ensemble group participants.

We explored three possible individual-difference factors that have been found to influence other risk judgments and decisions: numeracy (Fagerlin et al., 2007; Tak et al., 2015), graph literacy (Garcia-Retamero & Cokely, 2017; Okan et al., 2012), and decision fluency (Alter & Oppenheimer, 2009; Sanchez & Dunning, 2020).

Method

Recruitment

The survey was distributed through Dynata, a survey company widely used for COVID-19-related studies (Kobayashi et al., 2021; Lazarus et al., 2021; Lu et al., 2021). Dynata invited U.S. participants with prevalidated identities in its online consumer panel to participate. Those who agreed received a Qualtrics link. Dynata monitored recruitment to create samples matching the latest U.S. Census percentages for age, gender, and region (see Supplemental Table 3.1). Our preregistered power analysis estimated that 300 participants (100 per group) would detect a medium effect ($f = 0.25$) for a one-way analysis of variance (ANOVA) test, with a power of 0.95. The survey was conducted from April 20–28, 2021, when U.S. COVID-19 cases and deaths were decreasing and vaccination was becoming widely available. Five hundred seventy-one individuals used the link to view the study; 309 completed the full survey and were used in the data analysis. There were 104, 94, and 101 in the suite, ensemble, and combined groups, respectively. Carnegie Mellon University's Institutional Review Board approved all procedures. All participants provided informed consent.

Experiment Protocol

As seen in Figure 2, participants first read a short introduction describing the study's purpose: learning how they use multiple forecasts of COVID-19 cases. They then read a short decision scenario asking them to help hypothetical friends interpret COVID-19 forecasts and decide whether to send their child back to school, given the forecasts.

The scenario stated that their friends' child had not been vaccinated, that only a few schoolteachers were vaccinated, that their friends were worried about the new strain spreading in the country, and that their friends' decision threshold was 300 new cases per 100,000 people. They were then randomly assigned to one of the three groups. Ensemble and suite group participants completed the tasks twice, while combined group participants completed them once.

The graphs in Figure 3 reflect county-level forecasts for COVID-19 cases number per 100,000, to align with the decision scenario. We selected forecasts from 10 teams in CDC's display and created an ensemble using CDC's median value procedure. Each forecast has a 95% confidence interval (light color), a 50% confidence interval (darker color), and a best-guess (black line). Participants received instructions on reading the graphs. Suite and ensemble participants completed the dependent measures in the same order for both rounds. All participants completed the same final survey.

Data Analysis Approach

The statistical analysis plan was preregistered in Open Science Framework (Yang & Wong-Parodi, 2024). Hypotheses 1–3 on initial judgments and decisions were not preregistered, hence are exploratory. Supplemental Material Section 2 describes minor deviations from the preregistered plan. To assess whether participants were successfully randomized to three groups, we used one-way analysis of variance (ANOVAs) to compare the three groups for participants' age and a chi-square test for their gender. As preregistered, we identified participants who completed the study in 2 *SD*s less time than the group meantime and who failed a validity check. We included all participants in the reported analyses.

Predictions were elicited as (a) a 95% confidence interval for the next week's COVID-19 case number, using text-entry boxes for lower and upper bounds with the display range [0;1,500]; (b) the most likely range, with the options of less than 300, 300–600, 600–900, 900–1,200, or more than 1,200; and (c) whether the number was more likely to be above or below 300. The interval midpoints and widths were calculated using the upper and lower bounds that participants entered. Participants rated their confidence in how well they had interpreted the graph and in their decision advice, as well as the helpfulness of the graph for giving advice, on Likert scales from 1 = *not at all* to 5 = *extremely*. For these continuous variables, the groups were compared using nonparametric Kruskal–Wallis rank-sum tests and Welch one-way ANOVAs (See Supplemental Table 1.1).

Decision advice was elicited as 1 = *yes*, they should send the child to school and 0 = *no*, they should not. It was asked after reminding participants of their friends' decision threshold (300 cases). We used chi-square tests to compare advice across conditions.

To test for within-subject differences for repeated measures, we used generalized estimating equations (GEE) for the 95% confidence interval estimates. GEE estimates mean effects for a unit difference, allowing flexible specification of the correlation structure for within-unit data (Hubbard et al., 2010). For continuous variables of confidence and display evaluations, we used two-way ANOVAs with repeated measures. For discrete variables, we used repeated-measure logistic regression.

Participants' subjective numeracy was assessed with the four-item subjective ability subscale of the subjective numeracy scale (SNS; Fagerlin et al., 2007; Peters et al., 2019). We used the mean rating (Cronbach's $\alpha = .93$).

The subjective visual literacy was based on a three-item scale (Yang et al., 2023) that asked how well participants could use three common displays (graph, map, and manuals; Tversky et al., 2000), with 1 = *not at all* and 6 = *extremely well*. We used the mean rating (Cronbach's $\alpha = .79$).

Other individual measures included relevance of the hypothetical scenario to their own lives (1 = *not at all relevant* and 5 = *extremely*) and information-seeking frequency (1 = *never* and 5 = *very often*).

For exploring associations between participants' features and judgments, we used ordinary least squares regressions (controlling for demographics) and the Kendall τ correlation.

Results

Participants

Participants reported being, on average, 48.1 years old (*Mdn* = 48, *SD* = 16.8), with 36.9% female, 80.6% White, non-Hispanic, 67.6% having a bachelor's degree or higher, and 71.2% having a yearly household income above $50,000. On average, they reported being ideologically moderate (*M* = 4.0, *SD* = 1.9), where 1 = *extremely liberal* and 7 = *extremely conservative*. About 25.9% reported health care work experience. 16.5% reported having been infected with COVID-19, and 58.3% reported having been vaccinated against COVID-19.

Validity Check

There was no significant difference across the three groups in participants' age, $F(2, 306) = 1.1$, $p > .05$, or gender ($\chi^2 = 3.0$, $p > .5$), indicating successful randomization. To test comprehension and attention, we examined the consistency of participants' confidence intervals. Fifty-nine respondents (19.1%) had at least one inappropriate answer (e.g., estimating the expected number of case as < 300, with the most likely range of [900 to 1,200]). As mentioned, we included all responses. Repeating analyses without these individuals revealed similar patterns.

Initial Round Responses

Table 1 shows summary statistics.

> *Hypothesis 1:* As predicted, judgments of the initial interval midpoint were higher for the suite and combined groups, consistent with the higher 95% upper bound in the figure. However, the three groups were not statistically significant (Kruskal–Wallis: $\chi^2 = 1.6$, $p > .05$) nor were the pairwise differences (post hoc Dunn's tests; Supplemental Table 1.2).

> *Hypothesis 2:* Kruskal–Wallis tests found no significant difference in the three groups' initial confidence intervals (Kruskal–Wallis: $\chi^2 = 0.14$, $p > .05$). However, both one-way ANOVA, $F(2, 306) = 4.5$, $p = .012$, and Welch one-way ANOVA, $F(2, 194.7) = 4.1$, $p = .018$, revealed a significant difference. A post hoc Tukey test found that, as predicted, suite participants had significantly larger confidence intervals than did ensemble participants (diff = −157.8, $p = .01$).

Table 1
Judgments and Decisions, Mean, and Standard Deviation for Study 1

Condition	95% CI lower bound	95% CI upper bound	Interval width	Interval midpoint	More likely range (=above 300)	Decision = yes (go to the school)
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
First round						
Suite	191.9 (255.9)	578.1 (569.9)	413.5 (488.2)	385.0 (360.6)	82 (78.8%)	39 (37.5%)
Ensemble	205.3 (169.8)	456.8 (369.0)	255.7 (282.4)	331.0 (249.1)	70 (67.3%)	51 (49.0%)
Combined	207.5 (246.4)	480.2 (485.6)	304.8 (368.5)	343.8 (331.1)	65 (64.4%)	56 (55.4%)
Second round						
Suite	193.8 (197.2)	468.6 (434.2)	295.5 (344.6)	331.2 (284.7)	82 (78.8%)	38 (36.5%)
Ensemble	184.8 (164.6)	578.5 (492.7)	402.8 (446.1)	381.6 (288.7)	70 (67.3%)	49 (47.1%)

Note. CI = confidence interval.

The influence of the high outliers on these results is seen in the 95% upper bound estimates being higher for the suite and combined groups, compared to the ensemble group. There was no significant difference in the 95% lower bound estimates, consistent with there being no low outliers ($\chi^2 = 2.8, p > .05$). Nonetheless, as seen in Figure 4, some suite participants may have been strongly influenced by the two high outlier forecasts (University of California, Santa Barbara, Karlen).

Despite their higher upper bounds and midpoints, the suite and combined groups were no more likely to predict an actual value above 300 ($\chi^2 = 5.8, p > .05$).

> *Hypothesis 3:* The three groups gave significantly different decision advice ($\chi^2 = 6.8, p = .033$), as expected. The suite group, which saw greater uncertainty (as predicted), was the most cautious and least likely to advise their friends to send their child to school (37.5% vs. 49.0% and 55.4% for the ensemble and combined groups, respectively).

Second-Round Responses and Updating Effects (for Ensemble and Suite Groups)

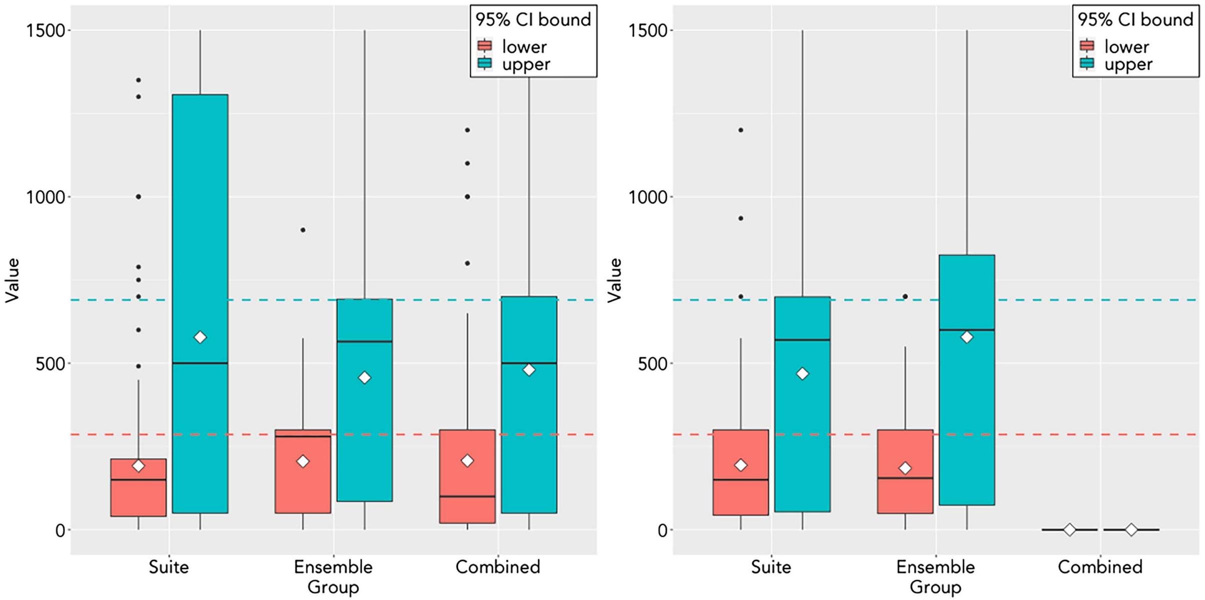
The bottom half of Table 1 and right-hand side of Figure 4 present suite and ensemble groups' responses, after seeing the added display. Table 2 presents GEE models comparing initial and updated responses.

The 95% upper bound is higher for the ensemble group ($M = 578.5, SD = 492.7$) than for the suite group ($M = 468.6, SD = 434.2$), although the difference is not statistically significant (Wilcoxon test: W = 4,813, $p > .05$). For both groups, the average interval widths are significantly narrower than in the initial round ($B = -117.9, p = .043$). However, they are now significantly wider for the ensemble group ($B = 265.0, p < .001$), reversing the pattern in the initial judgments. As a result, the intervals are now significantly narrower for the ensemble group than for the suite group ($B = -157.8, p < .01$). These differences are consistent with the outliers influencing revised ensemble judgments, as they influenced the initial suite judgments.

The interval midpoints remain similar in the two groups. There was no significant change in decision advice. See Supplemental Table 1.3.

Figure 4

Boxplot for Participants' 95% Confidence Interval of Predicted Cases for Study 1, With the Responses After the First-Round Display on the Left and After the Second-Round Display on the Right



Note. White diamonds denote the mean. The pink and blue horizontal dashed lines show the lower and upper bounds for the ensemble model prediction from CDC. CDC = Centers for Disease Control and Prevention; CI = confidence interval. See the online article for the color version of this figure.

Table 2

GEE Model Results for Interval Midpoints and Widths, With Repeated Measures for Display Rounds (Study 1)

	Interval midpoint				Interval width			
	1		2		1		2	
Variable	*B* (*SE*)	*p*	*B* (*SE*)	*p*	*B* (*SE*)	*p*	*B* (*SE*)	*p*
Ensemble group	−1.77 (29.24)	.95	−53.9 (42.8)	.21	−25.3 (39.4)	.52	−157.8 (55.0)	.0042
Round 2	−1.56 (29.24)	.96	−53.7 (44.8)	.23	14.6 (39.4)	.71	−117.9 (58.3)	.043
Ensemble Group × Round 2			104.4 (58.3)	.073			265.0 (77.8)	.00066

Note. GEE = generalized estimating equations; *SE* = standard error.

Responses to the Full Display

> *Hypothesis 4:* The combined group's initial judgments considered both displays, as did the suite and ensemble groups' second judgments. As predicted, one-way ANOVAs found no significant differences in the three groups' risk judgments: interval widths: $\chi^2 = 2.3$, $p > .05$; midpoints: $\chi^2 = 2.5$, $p > .05$; 95%CI lower bounds: $\chi^2 = 0.2$, $p > .05$; upper bounds: $\chi^2 = 2.7$, $p > .05$. Participants ratings were also similar for the three groups: confidence in their interpretation ($\chi^2 = 1.8$, $p > .05$), confidence in their decision advice ($\chi^2 = 1.8$, $p > .05$), and rated display helpfulness ($\chi^2 = 0.8$, $p > .05$).

Their final decisions were, however, significantly different ($\chi^2 = 7.4$, $p = .025$), with the suite group being more risk averse, as with its initial responses, contrary to Hypothesis 4, which predicted no order effects.

Display and Process Evaluations

Table 3 summarizes participants' evaluations, on the 5-point scale. All means were above the scale midpoint, indicating that, generally speaking, participants had moderate confidence in their ability to interpret the graphs and in their decision advice. They found the displays moderately helpful. There were no significant pairwise group differences. Two-way ANOVAs (with repeated measures for the display round) found no find significant group differences between initial and second ratings (for the suite and ensemble groups).

Respondent Features

We performed ordinary least squares regressions predicting participants' confidence interval widths, as a measure of uncertainty, from individual features. We looked at judgments for the full display so that all three groups were looking at the same information. Supplemental Material Section 4 has the full results. The width of participants' confidence intervals was unrelated to their subjective numeracy ($B = 1.8$, 95% CI [−50.8, 54.3], $p > .05$) or subjective visual literacy ($B = 1.7$, 95% CI [−58.8, 62.2], $p > .05$; Supplemental Table 4.1). Exploratory analyses found that participants with wider intervals rated the scenario as less relevant ($B = -35.2$, 95%CI [−66.9, −3.4], $p = .03$) and reported seeking COVID-19-related information less often ($\tau = -0.12$, $p < .01$).

Table 3

Process and Decision Evaluations, Mean, and Standard Deviation for Study 1

	Confidence in interpreting the graph	Confidence in decision	Perceived helpfulness of the graph
Condition	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
First round			
Suite	3.2 (1.2)	3.7 (1.1)	3.3 (1.3)
Ensemble	3.6 (1.2)	3.9 (1.1)	3.6 (1.3)
Combined	3.6 (1.2)	3.8 (1.1)	3.5 (1.2)
Second round			
Suite	3.4 (1.2)	3.7 (1.1)	3.4 (1.2)
Ensemble	3.5 (1.2)	3.9 (1.1)	3.5 (1.3)

Interval width was unrelated to scenario relevance or to having a medical background.

We used Kendall's τ for correlations among responses pooled across the three groups, given the skewed distributions. We highlight results that reflect on the validity of our measures (Supplemental Table 1.5). Respondents with higher interval midpoints gave wider intervals ($\tau = 0.73, p < .001$), were less confident in their interpretation of the displays ($\tau = -0.16$, $p < .001$), and rated the scenario as less relevant ($\tau = -0.13$, $p < .01$).

Participants were less likely to advise sending the child back to school if they rated the decision as more relevant personally (odds ratio = 1.4, $B = 0.3$, 95% CI [0.11, 0.5], $p < .01$).

Participants who rated the display as more helpful also rated themselves higher in numeracy ($\tau = 0.27, p < .001$) and visual literacy ($\tau = 0.32$, $p < .001$), and rated the scenario as more relevant ($\tau = 0.56$, $p < .001$).

Study 2

Design and Hypotheses

The results of Study 1 suggest that outliers drove differences between responses to these displays, as individual storm paths did for responses to hurricane forecasts. For the CDC COVID-19 forecasts, the outliers are unusually high values (Karlen, University of California, Santa Barbara), as the forecasts have similar lower bounds. If that is the case, then the visual salience of outliers should affect their impact. Following a suggestion from a reviewer, Study 2 examines this possibility by comparing responses to three versions of the suite display, varying whether it presents the individual forecasts in a random order (*random*), by increasing width of the prediction range (*range*), or by increasing the median estimate (*best-guess*). Participants were randomly assigned to receive one of these suite displays, after which they received the ensemble forecast (which was the same for all three groups). Figure 3 (Right pane) shows the three displays. As COVID-19 had waned by the time of the study, we replaced it with an influenza decision scenario, in which participants were asked to advise their hypothetical friends about vaccinating their child. Otherwise, we followed the same protocol and used the same forecasts as in Study 1.

We made the following hypotheses:

Initial Judgment and Decisions

> *Hypothesis 1:* Interval widths: We hypothesized that *range* group participants will give wider initial 95% confidence intervals for predicted new hospital admissions, compared to the other two groups.
>
> *Hypothesis 2:* Midpoints: We hypothesized that *best-guesses* group participants would give higher initial estimates for the midpoint of the 95% confidence interval compared to the other two groups.
>
> *Hypothesis 3:* Decision: We hypothesized that *random* group participants will be less likely to suggest vaccination to their friends compared to the other two groups.
>
> We also explored whether Study 1 results on updating and individual differences were replicated in Study 2, with suites of individual forecasts sorted by different visual features.

Method

Recruitment

The experiment was conducted on Prolific, an online platform widely used in behavioral decision research (Douglas et al., 2023). The survey company invited members of their participant pool living in the United States. It was conducted in late April 2024, with 106, 106, and 104 complete responses in the *random*, *range*, and *best-guesses* groups, respectively. Carnegie Mellon University's Institutional Review Board approved all procedures. All participants provided informed consent.

Data Analysis Approach

We preregistered the data analysis plan[1] which was the same as in Study 1, adapted to the differences in the two studies.

Results

Participants

Participants reported being, on average, 36.7 years old (*Mdn* = 34.5, *SD* = 11.6), with 48.1% female, 69.0% White, non-Hispanic, 57.9% having

[1] https://osf.io/cwnme/?view_only=335db739affa45e0917b3d6f1e0d7d85.

a bachelor's degree or higher, and 68.4% having a yearly household income above \$50,000. On average, they reported being politically liberal (M = 3.14, SD = 1.67), where 1 = *extremely liberal* and 7 = *extremely conservative*. Sixty (19.0%) reported having health care work experience. 38.3% reported having gotten a flu shot in the 2023–2024 season. They rated their relatively low frequency of getting flu shots as moderate (M = 2.75, with 1 = *never* to 5 = *often*) and our decision scenario as moderately relevant (M = 2.74, with 1 = *not at all relevant* and 5 = *extremely relative*). There was no significant difference across three groups in participants' age, $F(2, 313) = 0.47$, $p > .05$, or gender ($\chi^2 = 1.20$, $p > .05$), suggesting successful randomization.

Initial Round Responses

Figure 5 shows the upper and lower bound estimates given by participants in each group after reading the two rounds of graphs. As predicted in Hypothesis 1, participants in the *range* group gave wider initial 95% confidence intervals (M = 839.0, SD = 428.9) for predicted new hospital admissions than did participants in the *random* group (M = 657.2, SD = 400.0) or the *best-guesses* group (M = 712.4, SD = 395.2). Kruskal–Wallis test revealed a significant difference (Kruskal–Wallis: $\chi^2 = 7.24$, $p = .027$). Kruskal–Wallis tests also found significant difference in the three groups' initial midpoint estimates (Kruskal–Wallis: $\chi^2 = 23.91$, $p < .001$). However, contrary to our prediction, participants who saw the forecasts sorted by prediction range gave the highest midpoint estimates (M = 660.3, SD = 199.9), compared to the *random* group (M = 583.2, SD = 186.6) or the *best-guesses* group (M = 595.0, SD = 158.1). Contrary to our prediction as well, there was no significant difference among the three groups in their initial decision ($\chi^2 = 4.46$, $p > .05$), with a majority advising flu vaccination. Participants rated the three graphs similarly helpful and had similar levels of confidence in their graph comprehension and decision.

Responses to the Full Display

Consistent with Study 1 results, there were no significant differences in judgments or decision after participants read the full display with CDC's ensemble. Participants also rated the three graphs as similarly helpful and had similar confidence in their graph comprehension and advice. As in Study 1, respondents who rated the scenario as less relevant gave wider intervals ($\tau = -0.088$, $p < .05$), but just for estimates made after reading the initial display (Table 4).

Discussion

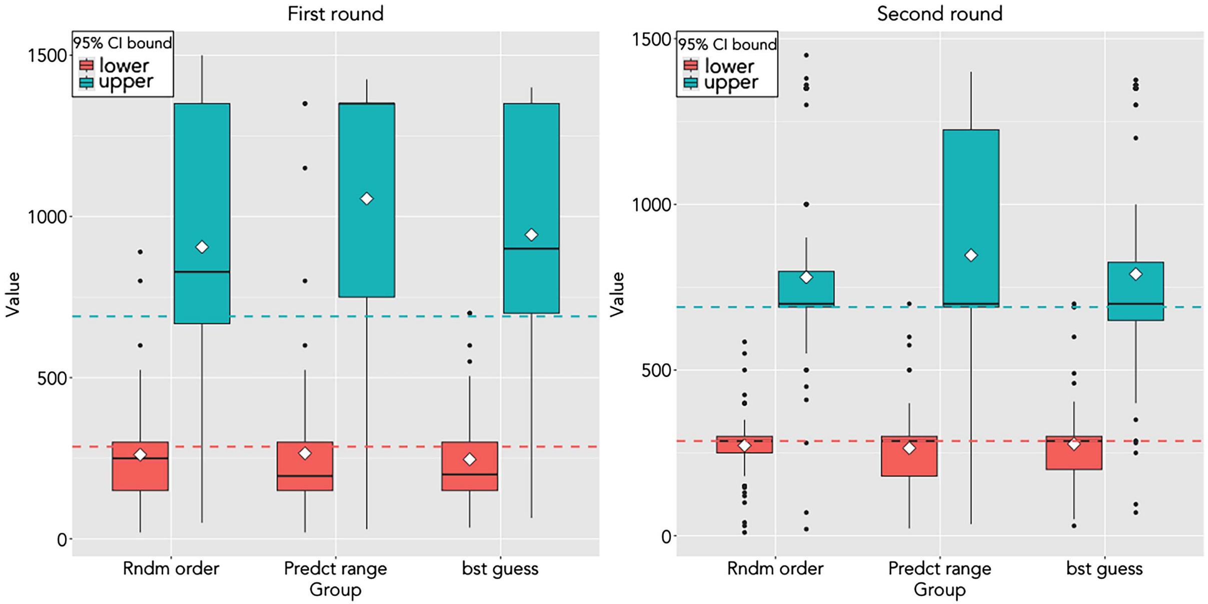
CDC, like many other organizations, routinely aggregates forecasts from multiple sources, each with its associated uncertainty. As an aid to users, CDC produces an ensemble summary to capture model uncertainty. We demonstrate an approach to evaluating the impact of choosing to display the suite or the ensemble or both and, if the latter, whether to present the two displays simultaneously or sequentially and in what order to sort the suite forecasts. We ask lay users to interpret the figures, apply their content to a hypothetical decision problem, and evaluate their experience. We used actual forecasts submitted to CDC by 10 competing teams. Participants were U.S. residents selected to match the population in age and gender from panels maintained by Dynata and Prolific. A large majority passed comprehension and attention checks, indicating that they could understand the displays. All were included in the analyses (whose conclusions were unaffected by excluding those who failed the tests).

In Study 1, the salient outliers led to wider confidence intervals for the initial judgments of the suite group than for the ensemble group. Consistent with these judgments, suite group members were less likely to recommend that their friends send their child to school (36.5%) compared to members of the ensemble and combined groups (49.0%, 55.4%). When the other display was added, the suite group's confidence intervals narrowed, while the ensemble group's widened—leaving them similar to one another and to those of the combined group, which saw both together. The midpoints of the groups' distributions showed the corresponding pattern. Thus, members of each group might have made different decisions, had they only seen one display, as might happen on a website that featured just one display, while offering the option of clicking through to the second display, for users who took that extra step. Although the suite and ensemble groups' judgments converged when the second display was added, the suite group's decisions remained more cautious.

Given the apparent influence of outliers in Study 1, Study 2 manipulated their salience in the

Figure 5

Boxplot for Participants' 95% Confidence Interval of Predicted Cases for Study 2, With the Responses After the First-Round Display on the Left and After the Second-Round Display on the Right



Note. White diamonds denote the mean. The pink and blue horizontal dashed lines were the lower and upper bounds for the Ensemble model prediction from CDC. CDC = Centers for Disease Control and Prevention; CI = confidence interval. See the online article for the color version of this figure.

Table 4

Judgments, Decisions, and Evaluations, Mean, and Standard Deviation for Study 2

Condition	95% CI Lower bound	95% CI Upper bound	Interval width	Interval midpoint	More likely range (=above 300)	Decision = yes
First round (suite)						
Random group	261.2 (149.8)	905.3 (368.6)	657.2 (400.0)	583.2 (186.6)	104 (98.1%)	97 (91.5%)
Range group	265.5 (219.4)	1055.2 (405.4)	839.0 (428.9)	660.3 (199.9)	101 (95.3%)	99 (93.4%)
Best-guesses group	246.7 (132.6)	943.3 (349.0)	712.4 (395.2)	595.0 (158.1)	103 (99.0%)	102 (98.1%)
Second round (combined)						
Random group	272.22 (88.6)	780.0 (278.9)	513.2 (307.2)	526.1 (133.6)	103 (97.2%)	100 (94.3%)
Range group	264.4 (107.8)	846.4 (333.8)	588.7 (369.4)	555.4 (159.5)	101 (95.3%)	101 (95.3%)
Best-guesses group	276.9 (112.6)	789.6 (319.2)	541.1 (347.4)	533.3 (139.8)	99 (95.2%)	101 (97.1%)

Note. CI = confidence interval.

display. It found that they had greater impact when forecasts were sorted by prediction range, which highlighted them. Here, too, judgments converged when participants saw the combined display.

The internal consistency of responses suggests construct validity of the measures. Participants with wider confidence intervals were less confident in their interpretation of the displays and rated the scenario as less relevant. Those who saw the scenario as more relevant also had narrower confidence intervals, lower midpoints, and greater updating.

Participants in all groups found the displays moderately helpful and were moderately confident in their ability to extract information from and base advice on them in both Study 1 and 2. These evaluations were more positive for participants who rated themselves as more numerate and more visually literate. Otherwise, exploratory analyses, based on the pooled final judgments of all groups, found that subjective numeracy and visual literacy were largely unrelated to any judgment.

The practical advice following from these results is to make the combined (suite and ensemble) display the default, in order to ensure that users see the full picture. Doing so would avoid the order effects seen with the suite and ensemble groups, where initial judgments differed from final ones, for upper bounds, confidence interval widths, and decision advice. Similar final evaluations with the different displays suggest that the added information in the combined display compensates for its added complexity. Showing the combined display also eliminated differences in judgments when the suite of individual forecasts was presented in different orders on the same page, in Study 2.

These results address a previously unstudied topic: how people interpret uncertainty within and across sources, simultaneously and sequentially, in visual displays. They reveal generally orderly responses, with individual differences related to engagement, but not to self-rated competencies.

The generalizability of these results is limited by its convenience sample, whose diversity suggests that the cognitive processes observed here might be generalized to people willing to participate in such studies. It is also limited by its suite 10 forecasts, taken from CDC, using its median aggregation method for generating the ensemble, and applied to hypothetical decisions reflecting friends' risk criteria. Padilla et al. (2022) found that suites with six or more individual forecasts evoked greater risk estimates than did ensemble displays. In qualitative analyses, Greis et al. (2017) found that some participants reported enjoying more detailed weather forecasts and the opportunity to compare them, whereas others did not need or want that much information.

Public service organizations are increasingly using suites of predictions and graphical displays to communicate uncertainty in weather, diseases, flooding (Stephens et al., 2012), climate change (Bruine de Bruin et al., 2024; Kause et al., 2020), and economic issues. Participants here were able to understand such forecasts, whether presented simultaneously or sequentially, and apply them to a hypothetical decision. They also found the displays moderately helpful and were moderately confident in their ability to interpret and apply them. The order effects indicate the value of presenting the combined display despite its greater complexity. If borne out with a more representative sample and additional test cases, such results may encourage scientists to display richer uncertainty information, including ensembles and summaries, without fear of scaring or confusing users (van der Bles et al., 2019, 2020).

References

Aikman, D., Barrett, P., Kapadia, S., King, M., Proudman, J., Taylor, T., De Weymarn, I., & Yates, T. (2011). Uncertainty in macroeconomic policy-making: Art or science? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4798–4817. https://doi.org/10.1098/rsta.2011.0299

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Boone, A. P., Gunalp, P., & Hegarty, M. (2018). Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of Experimental Psychology: Applied*, *24*(3), 275–295. https://doi.org/10.1037/xap0000166

Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, *88*(5), 651–668. https://doi.org/10.1175/BAMS-88-5-651

Bruine de Bruin, W., Rabinovich, L., Weber, K., Babboni, M., Ignon, L., Wald, R., Dean, M., Kashdan, A., & Luz, S. (2024). Improving figures for climate change communications: Insights from interviews with international policymakers and practitioners. *Climatic Change*, *177*(4), Article 57. https://doi.org/10.1007/s10584-024-03704-7

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., … Reich, N. G. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(15), Article e2113561119. https://doi.org/10.1073/pnas.2113561119

Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., & Ramos, M. (2007). Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*, *7*(2), 115–127. https://doi.org/10.1016/j.envhaz.2007.05.001

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), Article e0279720. https://doi.org/10.1371/journal.pone.0279720

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Gaba, A., Tsetlin, I., & Winkler, R. L. (2017). Combining interval forecasts. *Decision Analysis*, *14*(1), 1–20. https://doi.org/10.1287/deca.2016.0340

Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors*, *59*(4), 582–627. https://doi.org/10.1177/0018720817690634

Greis, M., Avci, E., Schmidt, A., & Machulla, T. (2017). Increasing users' confidence in uncertain data by aggregating data from multiple sources. *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 828–840). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025998

Grushka-Cockayne, Y., & Jose, V. R. R. (2020). Combining prediction intervals in the M4 competition. *International Journal of Forecasting*, *36*(1), 178–185. https://doi.org/10.1016/j.ijforecast.2019.04.015

Harries, C., Yaniv, I., & Harvey, N. (2004). Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, *17*(5), 333–348. https://doi.org/10.1002/bdm.474

Huang, S., Golman, R., & Broomell, S. B. (2024). Combining the aggregated forecasts: An efficient method for improving accuracy by stacking multiple weighting models. *Decision*, *11*(4), 668–683. https://doi.org/10.1037/dec0000245

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467–474. https://doi.org/10.1097/EDE.0b013e3181caeb90

Hullman, J., Kosara, R., & Lam, H. (2017). Finding a clear path: Structuring strategies for visualization sequences. *Computer Graphics Forum*, *36*(3), 365–375. https://doi.org/10.1111/cgf.13194

Kause, A., Bruine de Bruin, W., Fung, F., Taylor, A., & Lowe, J. (2020). Visualizations of projected rainfall change in the United Kingdom: An interview study about user perceptions. *Sustainability*, *12*(7), Article 2955. https://doi.org/10.3390/su12072955

Kobayashi, L. C., O'Shea, B. Q., Kler, J. S., Nishimura, R., Palavicino-Maggio, C. B., Eastman, M. R., Vinson, Y. R., & Finlay, J. M. (2021). Cohort profile: The COVID-19 Coping Study, a longitudinal mixed-methods study of middle-aged and older adults' mental health and well-being during the COVID-19 pandemic in the USA. *BMJ Open*, *11*(2), Article e044965. https://doi.org/10.1136/bmjopen-2020-044965

Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., Kimball, S., & El-Mohandes, A. (2021). A global survey of potential acceptance of a COVID-19 vaccine. *Nature Medicine*, *27*(2), 225–228. https://doi.org/10.1038/s41591-020-1124-9

Lu, J. G., Jin, P., & English, A. S. (2021). Collectivism predicts mask use during COVID-19. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(23), Article e2021793118. https://doi.org/10.1073/pnas.2021793118

Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology. Social judgment and decision making* (pp. 227–242). Psychology Press.

Ocean, N. (2024). Weighting ratings: Are people adjusting for bias in extreme reviews? *Journal of Experimental Psychology: Applied*, *30*(2), 391–409. https://doi.org/10.1037/xap0000497

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making*, *25*(4), 390–401. https://doi.org/10.1002/bdm.751

Padilla, L., Ruginski, I. T., & Creem-Regehr, S. H. (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, *2*(1), Article 40. https://doi.org/10.1186/s41235-017-0076-1

Padilla, L., Hosseinpour, H., Fygenson, R., Howell, J., Chunara, R., & Bertini, E. (2022). Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Scientific Reports*, *12*(1), Article 2014. https://doi.org/10.1038/s41598-022-05353-1

Park, S., & Budescu, D. V. (2015). Aggregating multiple probability intervals to improve calibration. *Judgment and Decision Making*, *10*(2), 130–143. https://doi.org/10.1017/S1930297500003910

Peters, E., Tompkins, M. K., Knoll, M. A. Z., Ardoin, S. P., Shoots-Reinhard, B., & Meara, A. S. (2019). Despite high objective numeracy, lower numeric confidence relates to worse financial and medical outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(39), 19386–19391. https://doi.org/10.1073/pnas.1903126116

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Lawrence Erlbaum Associates.

Poulton, E. C., & Freeman, P. R. (1966). Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin*, *66*(1), 1–8. https://doi.org/10.1037/h0023427

Ruginski, I. T., Boone, A. P., Padilla, L. M., Liu, L., Heydari, N., Kramer, H. S., Hegarty, M., Thompson, W. B., House, D. H., & Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition and Computation*, *16*(2), 154–172. https://doi.org/10.1080/13875868.2015.1137577

Sanchez, C., & Dunning, D. (2020). Decision fluency and overconfidence among beginners. *Decision*, *7*(3), 225–237. https://doi.org/10.1037/dec0000122

Stephens, E. M., Edwards, T. L., & Demeritt, D. (2012). Communicating probabilistic information from climate model ensembles-lessons from numerical weather prediction. *Wiley Interdisciplinary Reviews: Climate Change*, *3*(5), 409–426. https://doi.org/10.1002/wcc.187

Stephens, E. M., Spiegelhalter, D. J., Mylne, K., & Harrison, M. (2019). The Met Office Weather Game: Investigating how different methods for presenting probabilistic weather forecasts influence decision-making. *Geoscience Communication Discussions*, *2*(2), 101–116. https://doi.org/10.5194/gc-2-101-2019

Tak, S., Toet, A., & van Erp, J. (2015). Public understanding of visual representations of uncertainty in temperature forecasts. *Journal of Cognitive Engineering and Decision Making*, *9*(3), 241–262. https://doi.org/10.1177/1555343415591275

Toet, A., van Erp, J. B., Boertjes, E. M., & van Buuren, S. (2019). Graphical uncertainty representations for ensemble predictions. *Information Visualization*, *18*(4), 373–383. https://doi.org/10.1177/1473871618807121

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses and arrows: Diagrammatic communication with schematic figures. *International Conference on Theory and Application of Diagrams*, *1889*, 221–230. https://doi.org/10.1007/3-540-44590-0_21

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*(5), Article 181870. https://doi.org/10.1098/rsos.181870

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

Witt, J. K., & Clegg, B. A. (2022). Dynamic ensemble visualizations to support understanding for uncertain trajectories. *Journal of Experimental Psychology: Applied*, *28*(3), 451–467. https://doi.org/10.1037/xap0000370

Woloshin, S., Yang, Y., & Fischhoff, B. (2023). Communicating health information with visual displays. *Nature Medicine*, *29*(5), 1085–1091. https://doi.org/10.1038/s41591-023-02328-1

Wong-Parodi, G., & Strauss, B. H. (2014). Team science for science communication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(Suppl. 4), 13658–13663. https://doi.org/10.1073/pnas.1320021111

Yang, Y., & Wong-Parodi, G. (2024, May 5). *Communicating multiple COVID-19 forecasts*. https://osf.io/cwnme/?view_only=335db739affa45e0917b3d6f1e0d7d85

Yang, Y., Wong-Parodi, G., & Fischhoff, B. (2023). Visual displays for communicating scientific uncertainty in influenza forecasts. *Frontiers in Communication*, *8*, Article 1232156. https://doi.org/10.3389/fcomm.2023.1232156

Yaniv, I. (2004a). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I. (2004b). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. https://doi.org/10.1016/j.obhdp.2006.05.006

Received July 29, 2023
Revision received December 31, 2024
Accepted January 5, 2025 ■