AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General


ISSN: 0096-3445

https://doi.org/10.1037/xge0001388

# Are People More or Less Likely to Follow Advice That Is Accompanied by a Confidence Interval?

Celia Gaertig[1] and Joseph P. Simmons[2]
[1] Haas School of Business, University of California, Berkeley
[2] The Wharton School, University of Pennsylvania

Are people more or less likely to follow numerical advice that communicates uncertainty in the form of a confidence interval? Prior research offers competing predictions. Although some research suggests that people are more likely to follow the advice of more confident advisors, other research suggests that people may be more likely to trust advisors who communicate uncertainty. Participants ($N = 17{,}615$) in 12 incentivized studies predicted the outcomes of upcoming sporting events, the preferences of other survey responders, or the number of deaths due to COVID-19 by a future date. We then provided participants with an advisor's best guess and manipulated whether or not that best guess was accompanied by a confidence interval. In all but one study, we found that participants were either directionally or significantly *more* likely to choose the advisor's forecast (over their own) when the advice was accompanied by a confidence interval. These results were consistent across different measures of advice following and did not depend on the width of the confidence interval (75% or 95%), advice quality, or on whether people had information about the advisor's past performance. These results suggest that advisors may be more persuasive if they provide reasonably-sized confidence intervals around their numerical estimates.

***Public Significance Statement***
In 12 incentivized studies ($N = 17{,}615$), participants received numerical advice that was either accompanied by a confidence interval or not, and we observed how much they followed the advice. In all but one of those studies, we found that participants were either directionally or significantly more likely to choose the advisor's forecast (over their own) when the advice was accompanied by a confidence interval. Contrary to claims that people dislike expressions or acknowledgments of uncertainty, these results suggest that advisors may be more persuasive if they put confidence intervals around their numerical estimates.

*Keywords:* advice, uncertainty, overconfidence, confidence intervals, open science

*Supplemental materials:* https://doi.org/10.1037/xge0001388.supp



Many consequential decisions hinge on forecasts of uncertain events. An investment decision may hinge on forecasts of how stock prices might change in the near future, a gambling decision may hinge on forecasts of a sporting event, a decision to take a job or buy a house or have kids may hinge on forecasts about the course of a pandemic, etc. When making such forecasts, people often seek advice.

The future is uncertain, and advisors have to decide whether to communicate that uncertainty when they provide advice. For example, an advisor could merely provide a point estimate (e.g., "My best guess is that the stock price will increase by 5% this year") or s/he could communicate the uncertainty around that estimate by also providing a confidence interval (e.g., "My best guess is that the stock price will increase by 5% this year, and I am 95% sure it will increase by some value between 1% and 9%").

In this article, we attempt to answer a seemingly simple question. Are people more likely to choose an advisor's forecast over their own

Celia Gaertig https://orcid.org/0000-0002-5444-4999

We are grateful for financial support from the Bakar Faculty Fellowship at UC Berkeley, the Beatrice Foods Co. Faculty Research Fund at the University of Chicago, and the Wharton Behavioral Lab and the Baker Center Research Grant at the University of Pennsylvania. We thank Soaham Bharti, David Eskenazi, Beidi Hu, and Jordyn Schor for excellent research assistance. This work has previously been presented at the following conferences: ACR, SCP, SJDM, SPSP, and SPUDM. All of our studies were pre-registered. Our pre-registrations, as well as the data, code, and materials for all studies can be found here: https://researchbox.org/357.

Celia Gaertig and Joseph P. Simmons worked together to design the studies and to write the paper. Celia Gaertig served as the lead for the study execution and data analysis.

Correspondence concerning this article should be addressed to Celia Gaertig, Haas School of Business, University of California, Berkeley, CA, 94720, United States. Email: celia.gaertig@haas.berkeley.edu

when the advisor provides only a point estimate or when the advisor also provides a confidence interval around that estimate? In other words, are advisees more or less likely to follow the advice of an advisor who conveys uncertainty in the form of confidence intervals?

On this question the literature is mixed. On the one hand, advisors tend to provide overconfident advice. Undoubtedly, this at least partially reflects advisors' own overconfidence, as there is no reason to believe that advisors would be immune from the general tendency for people to believe that their own estimates are more accurate than they actually are, and to therefore provide confidence intervals that are overly precise (Alpert & Raiffa, 1982; Klayman et al., 1999; Moore et al., 2016; Moore & Healy, 2008; Soll & Klayman, 2004). But recent research suggests that advisors may also knowingly and strategically offer overconfident advice because they believe that there are benefits to doing so and/or costs to communicating uncertainty. And, indeed, that belief has some support in the published literature (Anderson et al., 2012; Price & Stone, 2004; Radzevick & Moore, 2011; Van Zant, 2022). For example, research suggests that advisees are more likely to choose advisors who provide more precise estimates (Radzevick & Moore, 2011). Moreover, research on the "confidence heuristic" suggests that people often presume that those who display more confidence have more knowledge (Price & Stone, 2004). As a result, they are more likely to prefer overconfident advisors to well-calibrated ones, at least when accuracy feedback is not easily available (Sah et al., 2013; Tenney et al., 2007, 2008).

The notion that overconfident advice may be more persuasive is also in line with the widespread belief that people dislike uncertainty (e.g., Kahneman, 2011; Tetlock & Gardner, 2015), a belief that persists even in the context of science communication (National Academies of Sciences, Engineering, and Medicine, 2017). In general, research suggests that overconfident statements of belief are driven at least in part by strategic, self-presentational concerns, and that those concerns may have some basis in reality (Anderson et al., 2012; Van Zant, 2022). From this, we might predict that advisees would be *less* likely to heed the advice of those who express uncertainty by putting confidence intervals around their estimates.

On the other hand, recent research has found that advisors are *not* trusted less, and sometimes even trusted *more*, when they incorporate uncertainty into their forecasts (Gaertig & Simmons, 2018; Gustafson & Rice, 2019; Howe et al., 2019; Joslyn & LeClerc, 2012, 2016; van der Bles et al., 2020). For example, Gaertig and Simmons (2018) found, across different domains, that advisors may be judged more positively when their forecasts communicate uncertainty, by, for example, providing a reasonably wide range of outcomes rather than a point estimate when predicting the outcomes of sporting events. Joslyn and LeClerc (2012, 2016) and Howe et al. (2019) found similar effects in their investigations of scientists' communications about climate projections. For example, Joslyn and LeClerc (2016) found that people were more trusting of scientists' climate projections, such as temperature forecasts or projections about rising sea levels, when the scientists' estimates were accompanied by a 90% confidence interval. And van der Bles et al. (2020) found no evidence that people exhibit diminished trust in a scientific source when the source provided a numerical range of outcomes or a confidence interval around a point estimate. This past research focused on evaluations of the advisor or the quality of the advisee's decision; it did not directly assess the extent to which people actually follow the advisor's advice.

The findings from this past research suggest that advisees may not be less likely to follow advice that is accompanied by a confidence interval. And, indeed, there are reasons to think that this form of uncertain advice may even be *more* persuasive. First, an advisor who pretends that s/he knows exactly what is going to happen next may not be very credible, at least to advisees who appreciate that the world is uncertain. Indeed, some advisors seem to recognize this. For example, in the context of predictions about the COVID-19 pandemic, the website FiveThirtyEight explained to their consumers, "COVID-19 models aren't made to be unquestioned oracles. They're not trying to tell us one precise future, but rather the range of possibilities given the facts on the ground" (Best & Boice, 2020). Second, advisees may use advisors' confidence intervals as a way to gauge whether their own estimate is reasonable—because it is inside the interval—or unreasonable—because it is outside the interval. As a consequence, advisees whose own estimates are outside of that interval may feel especially compelled to side with the advisor over their own potentially wayward estimate.

So which one is it? Are advisees more likely to follow advice that comes in the form of a single best guess, or when that guess is accompanied by a confidence interval?

## Research Overview

In 12 studies, we investigated whether people are more or less likely to follow advice when an advisor's best guess is accompanied by a confidence interval. Participants in our studies forecasted the outcomes of upcoming sporting events (Study 1–10), the number of deaths due to Covid-19 in the United States by a future date (Study 11), or the preferences of other survey responders (Study 12). In all studies, participants first made their own prediction. They then saw a forecast presented as coming from a statistical model (Studies 1–5 and 7–11) or a human advisor (Studies 6 and 12), and we manipulated whether the model's/advisor's forecast was or was not accompanied by a confidence interval.

We measured advice following in different ways across studies. In Studies 1, 2, 4–9, and 12, we asked participants to choose, for each of the items they were asked to predict, whether to submit their own forecast or the model's/advisor's forecast as their final, incentivized prediction. In Study 3, we instead asked participants to choose, all at once, between all of their own predictions or all of the model's forecasts. In Studies 10 and 11, we measured the extent to which participants adjusted their own prediction after seeing the model's prediction (i.e., a weight-of-advice measure).

While all of these design choices helped to establish the generalizability of our findings across different prediction tasks, different types of advisors, and different measures of advice following, we sought to further explore whether our results were robust to factors that might often differ across advice-giving contexts. Specifically, in some studies we introduced manipulations designed to test whether our findings differed depending on whether (1) people did or did not have access to how well the advisor had performed previously (Studies 4–6), (2) the quality of advice was high or low (Study 7), or (3) the confidence interval was wider (95%) or narrower (75%; Study 12).

The confidence intervals that we presented in our studies were based on real-world data and not on the subjective impressions of individual advisors (see Park & Budescu, 2015; Sniezek & van Swol, 2001). For the ten sports studies (Studies 1–10), we calculated the confidence intervals around advisors' best guesses using either a simple algorithm based on past data or odds set by professional oddsmakers. For Study 11, we obtained the confidence intervals from COVID-19 models featured on the website Fivethirtyeight.com at the time the study was run. And for Study 12, in which participants predicted the preferences of other survey responders, we generated the confidence intervals using pilot data and margin of error calculations for binary questions.

## Transparency and Openness

All of our studies were preregistered. All of our preregistrations, as well as our data, code, and materials, can be found on ResearchBox: https://researchbox.org/357. All studies were approved by the institutional review boards of either the University of Pennsylvania, University of Chicago, or University of California, Berkeley. Participants began each study by providing informed consent (via a Qualtrics survey).

# Studies 1–10

In Studies 1–10, participants predicted the outcomes of upcoming sporting events. These studies followed a similar procedure and so we describe them all at once.

## Method

### *Participants*

We conducted Studies 1–10 using U.S. participants from Amazon.com's Mechanical Turk (MTurk). We advertised Studies 1–3 as "a survey for National Basketball Association (NBA) basketball fans," Studies 4–7 as "a survey for Major League Baseball (MLB) baseball fans," and Studies 8–10 as "a survey for National Football League (NFL) fans." Participants received $1 for completing Studies 1, 3, 4, 5, and 8, $1.20 for completing Studies 2, 6, 7, and 10, and $1.50 for completing Study 9. Participants could earn up to an additional $0.90–$3.50 for accurate forecasting performance. In Studies 1–3, we decided in advance to recruit 400 participants, in Studies 4–7, we decided in advance to recruit 1,300 participants, and in Studies 8 and 10, we decided in advance to recruit 1,200 participants. Finally, in Study 9, we decided in advance to recruit 4,000 participants.

Our analyses included data from all participants who submitted a final prediction for at least one of the games but excluded those whom we preregistered to exclude. Across Studies 1–10, we preregistered different criteria for excluding participants on the basis of past participation, duplicate responses, extreme predictions (as defined for each sport in the study's preregistration; see the note for Table 1), and performance on an attention check (in Study 9 only). Table 1 summarizes our exclusions across Studies 1–10. After applying these exclusions, this left us with final samples of 381, 377, 376, 1,076, 1,222, 1,233, 1,205, 1,156, 3,842, and 1,166 participants in Studies 1–10, respectively. These samples averaged 33.9–37.9 years of age and were 33.6%–47.4% female.

### *Design*

In each of Studies 1–10, we provided participants with advice in the form of a model's or human advisor's best guess, and we manipulated between subjects whether this best guess was or was not accompanied by a confidence interval. In Studies 4–7, we additionally manipulated a second factor, which we describe below.

### *Procedure*

Table 2 provides all methodological details for Studies 1–10. The studies followed a similar procedure. In each study, participants were asked to predict the outcomes of a series of sports games on the day on which the games were played or a few days earlier. Participants in Studies 1–3 predicted NBA basketball games, participants in Studies 4–7 predicted MLB baseball games, and participants in Studies 8–10 predicted NFL football games.

For each study, we randomly selected a specific number of games that began no earlier than 7 p.m. Eastern Time on the selected game day (see Table 2 for the number of games used in each study). We posted the study in the morning of the game day or a few days earlier to ensure that data collection would be completed before the games started.[1] For each game, participants were presented with the game's start time, the names of the home and visiting teams, and information on how many wins and losses each of the teams had in the season so far. In addition, in Studies 1–3 and 8–10, participants were also shown how many points the teams had scored and allowed on average in each game thus far in the season, and, in Studies 4–7, participants were also shown the probable starting pitchers for each team. The games were presented in a random order and were presented one at a time on the screen.

For each game, participants were asked to make a prediction. In Studies 1–3, participants predicted how many points one of the teams would score in each NBA game, in Studies 4–7, participants predicted how many hits two teams would accumulate in each MLB game, and in Studies 8–10, participants predicted how many total points the two teams would score in each NFL game. We asked participants to make their prediction on the same screen on which we showed them the details for the game. Participants made their prediction by typing a number in a textbox.

After participants made their own prediction for every game in the study, we told them that they would next see the prediction that a statistical model (or a human advisor in Study 6) made for each of the games. Participants were presented again with each of the games, one at a time and in random order. For each game, we repeated the game details, the prediction question, and the participant's own prediction, and we also showed participants the model's/advisor's prediction. Figure 1 shows this screen for one of the games in Study 9.

**Confidence-Interval Manipulation.** In all 10 studies, we manipulated between subjects whether or not the model's/advisor's best guess was accompanied by a confidence interval. That is, in the no-confidence-interval condition, the prediction was presented as a

[1] Studies 5 and 7 were run over multiple game days to ensure that we could collect our target sample size. For these studies, we created multiple surveys, each with the games being played that day. We stopped data collection before the start of the games on one day and then resumed data collection the next day, asking different participants to predict a different set of games that were going to be played on that day.

**Table 1**
*Demographics and Number of Participants Excluded in Studies 1–10*

| | | | | | Gender | | | Preregistered exclusion criteria | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Study | Sport | Starting *n* | Final *n* | $M_{age}$ | Female (%) | Male (%) | Undisclosed (%) | IP address in the previous study | Duplicate IP address | Did not pass attention check | Made extreme predictions[a] |
| 1 | NBA | 401 | 381 | 34.9 | 33.6 | 65.6 | 0.8 | — | — | — | 20 |
| 2 | | 397 | 377 | 35.1 | 34.7 | 65.1 | 0.3 | — | — | — | 20 |
| 3 | | 399 | 376 | 34.5 | 35.5 | 64.3 | 0.3 | — | — | — | 23 |
| 4 | MLB | 1,292 | 1,076 | 35.1 | 39.3 | 60.2 | 0.6 | 13 | 32 | — | 171 |
| 5 | | 1,306 | 1,222 | 35.2 | 40.5 | 59.0 | 0.5 | 23 | 26 | — | 35 |
| 6[b] | | 1,314 | 1,233 | 33.9 | 43.1 | 56.6 | 0.3 | 37 | 15 | — | 29 |
| 7[b] | | 1,315 | 1,205 | 34.8 | 47.4 | 51.6 | 1.0 | 56 | 28 | — | 26 |
| 8[b] | NFL | 1,214 | 1,156 | 36.3 | 39.3 | 60.5 | 0.2 | — | 44 | — | 14 |
| 9[b] | | 4,342 | 3,842 | 37.9 | 44.6 | 54.8 | 0.6 | — | 161 | 312 | 27 |
| 10[b] | | 1,211 | 1,166 | 35.6 | 42.9 | 56.8 | 0.3 | 15 | 27 | — | 3 |

*Note.* Cells containing a dash indicate that we did not preregister to exclude participants on this basis, and therefore did not. NBA = National Basketball Association; MLB = Major League Baseball; NFL = National Football League; WOA = Weight of Advice.
[a]Extreme predictions were defined as follows in the different studies. In Studies 1–3 (NBA), we excluded participants who made at least one prediction greater than 160 points or less than 60 points. In Studies 4–7 (MLB), we excluded participants who made at least one prediction greater than 40 hits or at least two predictions less than 3 hits. And in Studies 8–10 (NFL), we excluded participants who made at least one prediction greater than 120 points.
[b]In addition (and as preregistered), in Studies 7–10, we excluded 806 (5.6%), 563 (3.5%), 2,589 (5.2%), and 610 (4.4%) observations, respectively, for which the participant's and the model's/advisor's predictions were identical. And, in Study 10, to calculate WOA, we also excluded 1,266 observations (9.1%) for which the participant's and the model's/advisor's prediction differed by exactly one.



best guess only (e.g., "The statistical model's best prediction is that the two teams will score 45 total points."). In the confidence-interval condition, we, of course, added a confidence interval to the best guess (e.g., "According to the statistical model, there is a 75% chance that the two teams will score between 30 and 60 total points. Its best prediction is that the two teams will score 45 total points."). That is, the only difference between the experimental conditions was whether or not we provided participants with a confidence interval around a model's/advisor's best guess. The best guess itself was held constant across conditions.

For each prediction question, we obtained a best guess from data-driven algorithms or well-calibrated online betting markets, and we created a confidence interval around that best guess based on past data. In Studies 1 and 8–10, we used a 75% confidence interval around the best guess ($\pm 12$ points in Study 1 and $\pm 15$ points in Studies 8–10), and in Studies 2–7, we used a 90% confidence interval around the best guess ($\pm 17$ points in Studies 2–3 and $\pm 9$ hits in Studies 4–7). To help establish the generalizability of our results, we used different confidence interval widths (i.e., 75% and 90%) across studies (see Table 2).

**Measuring Advice Following.** We measured advice following in different ways across the 10 studies (see Table 2).

***Choice for Each Game.*** In Studies 1, 2, and 4–9, participants were asked to choose, for each game separately, whether they would like to use their own prediction or the model's/advisor's prediction as their official prediction. They were told that their official prediction would be the one that determined their bonus payment. On the same screen on which they saw their own prediction and the model's/advisor's prediction, we asked them, "Which prediction would you like to submit as your official prediction?" Participants indicated their choice by selecting one of the two options. We measured advice following by calculating the percentage of observations for which participants chose the model's/advisor's best guess over their own prediction.

***One Choice for All Games.*** Instead of asking participants to make this choice for each game, in Study 3, we asked them to make this choice once for all of the games at the end of the study. That is, in this study, participants first saw the model's predictions for all games (along with their own predictions). Then, on a separate screen, we asked them to indicate whether they wanted to use the statistical model's predictions or their own predictions as their official predictions for all games. To facilitate this choice, we showed participants an overview table that listed all the games again along with the model's prediction and the participant's own prediction for each game. Participants indicated their choice by selecting one of two options: (1) "Use the statistical model's predictions to determine my bonus for all 7 games" or (2) "Use my own predictions to determine my bonus for all 7 games."

***Weight of Advice.*** Finally, in Study 10, instead of asking participants to choose between their own prediction and the model's prediction for each game, we measured the extent to which participants adjusted their final prediction after seeing the model's advice using a weight-of-advice (WOA) measure (e.g., Harvey & Fischer, 1997; Yaniv, 2004). On the same screen on which participants saw the model's prediction, we told them, "You now have a chance to change your prediction. Please make your final prediction." Participants made their final prediction by typing a number into a textbox, just like they did for their initial prediction. To calculate WOA, we used the following formula:

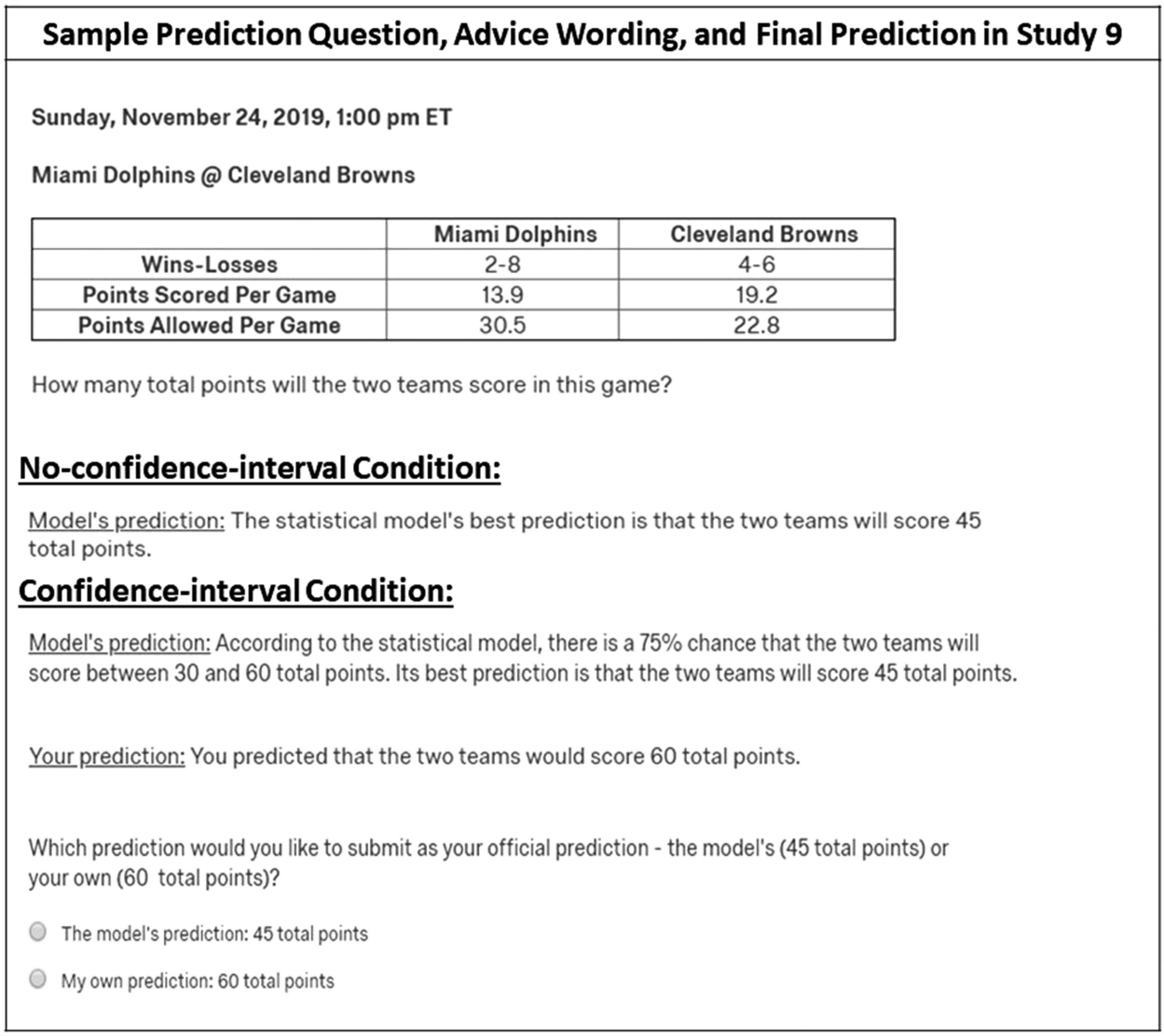
$$\text{WOA} = \frac{\text{Final Prediction} - \text{Initial Prediction}}{\text{Model's Prediction} - \text{Initial Prediction}}$$

**Table 2**
*Prediction Questions and Experimental Designs for Studies 1–12*

| Study | Final *n* | Domain | Prediction question | Incentive | Number of predictions | Advice from | Confidence interval | Measure of advice following | Information on past performance | Number of past items | Advice quality |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 381 | NBA | How many points will the (team name) score in this game? | $0.10 for each game prediction within 1 point of outcome | 9 | Model | 75% (±12 points) | Percent of time participant chose the model's best guess over their own prediction | No | N/A | Good |
| 2 | 377 | | | $0.25, $0.20, $0.15, or $0.10 for each game prediction within 0, 1, 3, or 5 points of outcome | 9 | Model | 90% (±17 points) | | Yes | 14 | Good |
| 3 | 376 | | | | 7 | | | Percent of participants choosing the model's best guess over their own prediction | | 11 | |
| 4 | 1,076 | MLB | How many hits will the two teams accumulate in this game? | $0.20, $0.10, or $0.05 for each game prediction within 0, 1, or 2 hits of outcome | 15 | Model | 90% (±9 hits) | Percent of time participant chose the model's/advisor's best guess over their own prediction | Manipulated | 15 | Good |
| 5 | 1,222 | | | | 15 per game day | Model | | | Manipulated | 15 | Good |
| 6 | 1,233 | | | | 13 | Human Advisor | | | Manipulated | 15 | Random |
| 7 | 1,205 | | | | 11 or 12 per game day | Model | | | Yes | 15 | Manipulated |
| 8 | 1,156 | NFL | How many total points will the two teams score in this game? | $0.25, $0.20, $0.15, or $0.10 for each game prediction within 0, 1, 3, or 5 points of outcome | 14 | Model | 75% (±15 points) | Percent of time participant chose the model's best guess over their own prediction | Yes | 14 | Good |
| 9 | 3,842 | | | | 13 | | | | Yes | 13 | Good |
| 10 | 1,166 | | | | 12 | | | Weight of advice for each game | Yes | 14 | Good |
| 11 | 3,622 | COVID-19 | What will the number of confirmed deaths due to COVID-19 in the United States be as reported at 7pm Eastern Time on August 1, 2020? | $0.25 if within 5% of correct answer | 1 | Model | 95% | Weight of advice | No | N/A | Good |
| 12 | 1,959 | Preferences | 10 preferences/behaviors of previous survey responders, e.g., "How many of the 100 people we surveyed said that they prefer politics to sports?" | $0.15, $0.10, or $0.05 for each prediction that was within 1, 3, or 5 of the correct answer | 10 | Human Advisor | Manipulated (75% vs. 95%) | Percent of time participant chose the advisor's best guess over their own prediction | No | N/A | Manipulated |

*Note.* In each of Studies 1–11, we manipulated whether participants were presented with only the model's/advisor's best guess (No-confidence-interval condition) or a best guess accompanied by a confidence interval (Confidence-interval condition). In Studies 4–7, we also manipulated an additional factor (see two rightmost columns), and hence these studies included four conditions in total. Finally, in Study 12, participants were randomly assigned to a No-confidence-interval condition, a 75%-confidence-interval condition, or a 95%-confidence-interval condition, and we also manipulated within subjects the difference between the best guess and the correct answer (i.e., the quality of the advice). NBA = National Basketball Association; MLB = Major League Baseball; NFL = National Football League.

**Figure 1**
*Sample Stimulus in Study 9*




Participants were randomly assigned to see only the model's best guess (no-confidence-interval condition) or to see the best guess accompanied by a confidence interval (confidence-interval condition).



WOA measures how much weight participants gave to the advice in making their final, incentivized prediction. If the final prediction was identical to the model's best guess, the advice was fully incorporated and WOA is 1. If the final prediction was identical to the initial prediction, the advice was ignored and WOA is 0. Values between 0 and 1 indicate that the advice was used to some extent, and negative values indicate that participants moved away from the advice. As preregistered, we winsorized the WOA measure at −1 and 1, meaning that observations for which WOA was greater than 1 or smaller than −1 were treated as 1 and −1, respectively (e.g., Harvey & Fischer, 1997; this was the case for 1,021 observations, 7.3% of the total sample).

**Additional Manipulations.** In addition to manipulating whether the advice was accompanied by a confidence interval, in some of the studies we also manipulated (a) whether or not we provided participants with information about the model's/advisor's past performance or (b) whether the advice that we gave participants was well-calibrated or random (see also Table 2).

***Information on the Model's/Advisor's Past Performance.*** In Studies 4–6, we examined whether the effect of the confidence-interval condition on advice following changed based on whether participants were provided with information about how the model/advisor had performed previously. Past research has found that people are less likely to accept advice from a statistical model when they have seen that model (inevitably) make mistakes (Dietvorst et al., 2015). We thought that perhaps confidence intervals would work similarly; they might add an air of credibility when participants lack information about how they perform, but be less compelling once

participants see them in action, especially after they (inevitably) observe that some outcomes fall outside of those confidence intervals.

In Studies 4–6, participants in the past-performance-information condition were first presented with a "training" block prior to making their predictions. This "training" block displayed the model's/advisor's predictions for a number of games that had recently been played, alongside the actual game outcomes. For example, in Study 5, participants were asked to make predictions for 15 MLB games being played on July 14th and 15th, and so we showed participants in the past-performance-information condition predictions made by the model for the 15 games played on the most recent game day prior to the study game days (July 9th), along with the outcome for each game. (Table 2 provides the number of games for which past performance was provided in each study.) Importantly, we presented the model's/advisor's predictions in the "training" block either with or without confidence intervals, depending on which condition participants were in. Thus, the forecasts that participants saw during the "training" block looked exactly like the advice they received later in the study. Participants in the no-past-performance-information condition did not receive any information about the model's/advisor's past performance prior to making their predictions.

In Studies 1–3 and 7–10, we did not manipulate whether or not participants received information about the model's past performance. Specifically, in Study 1, none of the participants saw how the model had performed previously, whereas in Studies 2, 3, and 7–10 all of the participants saw how the model had performed previously.

***Advice Quality.*** In most of our studies, we provided participants with good, well-calibrated advice. But obviously, advice can be either good or bad, and it is important to know whether the presence of confidence intervals influences advice following differently depending on the quality of the advice. In Study 7, we manipulated whether we provided participants with good, well-calibrated advice or with random advice. In this study, participants predicted the number of hits that would be accumulated in MLB games. Those in the good advice condition saw advice that was based on a simple model that we built to predict the number of hits that would occur in a baseball game. Those in the random-advice condition saw advice that was generated by randomly adding or subtracting a random number between 5 and 8 to or from the good advice. In the remaining studies, we did not manipulate the quality of the advice, but either provided participants with good advice (Studies 1–5 and 8–10) or with random advice (Study 6, which used a hit prediction randomly drawn from a uniform distribution ranging from 13 to 26 hits).

**Incentives.** Participants in every study could earn bonus money that was tied to the accuracy of their predictions, but we used different incentives across the studies. In Study 1 (NBA), participants received \$0.10 for every prediction that was within 1 point of the true outcome. In Studies 2–3 (NBA) and 8–10 (NFL), participants received \$0.25 for every prediction that they got exactly right, and they received \$0.20, \$0.15, or \$0.10 if the prediction was within 1, 3, or 5 points of the true outcome, respectively. And in Studies 4–7 (MLB), participants received \$0.20 for every prediction that they got exactly right, and they received \$0.10 or \$0.05 if the prediction was within 1 or 2 hits of the true outcome, respectively.

**Additional Information Provided in the MLB Studies.** In Studies 4–7, participants predicted how many hits two teams would accumulate in MLB games. In Study 4, we did not provide participants with any additional information about accumulated hits. Perhaps because of this, 14% of the sample made very extreme predictions, a fact we took as evidence that participants may not have had enough information about how to make these kinds of relatively unfamiliar predictions. Thus, we decided to provide participants in Studies 5–7 with some additional information before we asked them to make their own predictions. Specifically, before making their own predictions, we showed participants information about the number of hits accumulated in a series of previous MLB games. For example, in Study 5, participants were asked to predict hits accumulated in games played on July 14th and 15th, and so we first showed them how many hits were accumulated in the games played on July 9th. The purpose of this information was to familiarize participants with the typical range of the number of hits accumulated in MLB games. We presented participants with a table that listed the home and visiting teams that played each other that day, their win–loss records, their starting pitchers, and how many total hits they accumulated.

To ensure that participants attended to this information, we asked them three attention check questions. These questions were presented on the same screen as the information we provided. We asked participants to indicate how many hits were accumulated in one specific game from the table and to indicate the fewest and most amount of hits in any of the games. If participants answered any of these three questions incorrectly, they were asked to take another look at the table and to answer the attention check questions again. If participants answered any of the three questions incorrectly after three attempts, they were informed that they were ineligible to participate in the study.

**Exploratory Measures.** At the end of each study, participants were asked a series of exploratory questions. We provide a full list of these exploratory questions and their exact wording in the online supplemental material 2. In all studies, we asked participants about their motivation to make accurate predictions: "How hard did you try to make accurate predictions when making your own predictions about the game outcomes?" ($1 = I$ *did not try hard at all*; $7 = I$ *tried extremely hard*).

In Studies 2–10, we also asked participants a series of additional questions before this motivation question. First, we asked participants how much confidence they had in the model's/advisor's best predictions and in their own predictions: "How much confidence do you have in the model's/advisor's best predictions?" ($1 =$ *none*; $5 =$ *a lot*), and "How much confidence do you have in your own best predictions?" ($1 =$ *none*; $5 =$ *a lot*). Second, we asked participants to indicate how far off they thought the model's/advisor's best guesses and their own best guesses would be on average. For example, for those studies that asked participants to make point predictions (Studies 2, 3, and 8–10), we asked, "Considering the [number of games] games you were just asked to predict … on average, how many points do you think the model's best guesses will be away from the true point totals?" (dropdown list from "0" to "40 or more" in 1-point increments) and "… on average, how many points do you think your best guesses will be away from the true point totals?" (dropdown list from "0" to "40 or more" in 1-point increments).[2] Third, we also asked them to indicate the percentage of games for which they thought the model's/advisor's best prediction would be better than their own: "In general, for what percentage of games do you think the model's/advisor's

[2] In Studies 4 and 5, these "distance" questions contained a typo, such that they referred to "points" instead of "hits," even though these studies elicited hit predictions.

best prediction would be better than your own?" (dropdown list from "0%" to "100%" in 1% increments).

**Sports Knowledge.** At the end of each study, we presented participants with a set of questions designed to assess their knowledge about the sport they were predicting. In Studies 1–3 (NBA) and 4–7 (MLB), we asked participants to identify the teams of four different players and to identify which teams had the best and worst records at the time of the study. In Studies 8–10 (NFL), we asked participants, for a series of eight different pairs of teams, which of the teams had scored more points this season so far. Participants were asked to answer these questions without looking up the answers. We used participants' answers to these questions to construct a variable indicating how much participants knew about the sport they were predicting. We present the results from analyses that include this sports knowledge variable in the online supplemental material 3. In addition, in Study 10, we also asked participants to indicate how closely they followed the National Football League (1 = *not at all closely*; 7 = *extremely closely*) and to pick their favorite NFL team from the complete list of teams.

**Demographics.** Each study ended by asking participants to indicate their age in an open-ended textbox and their gender by selecting from the options "Male", "Female", and "Prefer not to disclose."

## Results

### Was It Beneficial to Follow the Advice?

Before presenting the main analysis, we first wanted to establish that it was in most cases wise for participants to follow the model's/advisor's advice. To test whether the forecasts in our studies outperformed participants' predictions, we computed, for each participant, how far the participant's predictions were from the outcomes of the games on average and how far the model's/advisor's predictions were from the outcomes of the games on average. We then conducted t-tests to compare the mean distance between participants' own predictions and the outcome and the mean distance between the model's/advisor's predictions and the outcome. The results showed that the model/advisor outperformed participants in every study/condition designed to produce good advice (i.e., in Studies 1–5 and 8–10, and the good advice condition in Study 7; all $ps < .001$; see Table S1 in the online supplemental materials).[3] That is, the model's/advisor's best guesses in these studies were on average closer to the true outcomes than participants' own guesses. Hence, it would have been better for participants to follow the advice rather than rely on their own prediction.

### Analysis Plan

We preregistered to analyze each study separately. In Studies 1–2 and 4–10, each participant contributed multiple observations to the dataset, one for each game they predicted. In Study 3, participants only made one choice for all games, and so each participant contributed only one observation to the dataset. In Studies 1–3 and 8–9, we regressed participants' choice of the model (1 = they chose the model; 0 = they did not choose the model) on the confidence-interval condition (1 = confidence interval; 0 = no confidence interval). And in Study 10, we regressed the WOA measure on the confidence-interval condition (1 = confidence interval; 0 = no confidence interval). Except for Study 3, all of these analyses included fixed effects for game and clustered standard errors across participants.

Studies 4–7 each included an additional manipulation other than the confidence-interval manipulation, and thus the regressions included an additional factor. In Studies 4–6, we regressed participants' choice of the model/advisor (1 = they chose the model/advisor; 0 = they did not choose the model/advisor) on (1) the confidence-interval condition (+0.5 = confidence interval; −0.5 = no confidence interval), (2) the training condition (+0.5 = training; −0.5 = no training), and (3) their interaction. And in Study 7, we regressed participants' choice of the model (1 = they chose the model; 0 = they did not choose the model) on (1) the confidence-interval condition (+0.5 = confidence interval; −0.5 = no confidence interval), (2) the advice quality condition (+0.5 = good advice; −0.5 = random advice), and (3) their interaction. All analyses included fixed effects for game and clustered standard errors across participants.

As preregistered, in Studies 7–9, we also dropped from the analysis those observations for which a participant's own prediction was identical to the model's best guess; and in Study 10, in order to generate the weight-of-advice measure, we dropped both those observations for which a participant's own prediction was identical to the model's best guess and those for which it differed by exactly one point. (See superscript [b] in the note for Table 1 for more detail about these exclusions.)

### Main Analysis

Were participants less likely to follow advice that came with a confidence interval? On the contrary, we found that participants were somewhat *more* likely to follow advice that was accompanied by a confidence interval. As shown in Table 3, this effect was directional in nine out of ten studies, and statistically significant in five of those nine. Overall, participants were about 3 percentage points more likely to choose an advisor's advice over their own when it was accompanied by a confidence interval; moreover, in Study 10 we found that providing confidence intervals around advisors' estimates increased the weight that participants gave to the advice by about 3–4 percentage points. Although effects of this magnitude usually require very large samples to detect (and, indeed, many of our studies were probably somewhat underpowered to detect them), they are similar in magnitude to many well-touted effects in psychology and behavioral economics, such as implementation intentions (e.g., Nickerson & Rogers, 2010) and other "nudges" (e.g., Della Vigna & Linos, 2022).

### Analysis of Training Information

In Studies 4–6, we manipulated whether or not we presented participants with information about the model's/advisor's past performance. In Study 4, participants were more likely to follow the advice when they saw how the model had performed previously ($b = 0.091$, $SE = 0.019$, $t = 4.88$, $p < .001$). Note that in this study, we did not provide any information about how many hits are typically scored in MLB games, and hence participants may have used the

[3] Participants were only able to outperform the model/advisor in studies in which the advice was random, and thus of poor quality. This was the case in Study 6 ($p < .001$), in which we generated the advice on hit predictions by randomly adding or subtracting a random number between 5 and 8 to or from the good advice. And it was also the case in the random advice condition in Study 7 ($p < .001$), in which we generated the advice by randomly selecting a number between 13 and 26 hits. See online supplemental material 1 for the full results across all studies.

**Table 3**
*Results for Studies 1–12*

| Study | Final *n* | Domain | Measure of advice following | No confidence interval | Confidence interval | Main effect of confidence interval (vs. no confidence interval) | Which was followed more? |
|---|---|---|---|---|---|---|---|
| 1 | 381 | NBA | Percent of time participant chose the model's best guess over their own prediction | 32.3% | 35.4% | $b = 0.031$, $SE = 0.031$, $t = 1.01$, $p = .314$ | Confidence Interval (ns) |
| 2 | 377 | | | 18.7% | **25.0%** | $b = 0.063$, $SE = 0.026$, $t = 2.45$, $p = .015$ | **Confidence Interval (significant)** |
| 3 | 376 | | Percent of participants choosing the model's best guesses over their own predictions | 10.4% | **17.5%** | $\chi^2 = 3.89$, $p = .048$ | **Confidence Interval (significant)** |
| 4 | 1,076 | MLB | Percent of time participant chose the model's/advisor's best guess over their own prediction | 50.6% | 49.6% | $b = -0.012$, $SE = 0.019$, $t = -0.65$, $p = .518$ | No Confidence Interval (ns) |
| 5 | 1,222 | | | 39.6% | 40.5% | $b = 0.010$, $SE = 0.016$, $t = 0.62$, $p = .536$ | Confidence Interval (ns) |
| 6 | 1,233 | | | 31.1% | **35.4%** | $b = 0.044$, $SE = 0.015$, $t = 2.91$, $p = .004$ | **Confidence Interval (significant)** |
| 7 | 1,205 | | | 30.1% | 32.4% | $b = 0.023$, $SE = 0.015$, $t = 1.57$, $p = .117$ | Confidence Interval (ns) |
| 8 | 1,156 | NFL | Percent of time participant chose the model's best guess over their own prediction | 35.3% | **39.1%** | $b = 0.037$, $SE = 0.019$, $t = 1.99$, $p = .047$ | **Confidence Interval (significant)** |
| 9 | 3,842 | | | 30.4% | **33.1%** | $b = 0.026$, $SE = 0.010$, $t = 2.75$, $p = .006$ | **Confidence Interval (significant)** |
| 10 | 1,166 | | Weight of advice for each game | 0.30 | 0.32 | $b = 0.021$, $SE = 0.016$, $t = 1.28$, $p = .201$ | Confidence Interval (ns) |
| 11 | 3,622 | COVID-19 | Weight of advice | 0.48 | **0.52** | $b = 0.038$, $SE = 0.014$, $t = 2.78$, $p = .006$ | **Confidence Interval (significant)** |
| 12 | 1,959 | Preferences | Percent of time participant chose the advisor's best guess over their own prediction | 34.8% | **40.6%** | $b = 0.058$, $SE = 0.014$, $t = 4.23$, $p < .001$ | **Confidence Interval (significant)** |
| | | | | | **40.8%** | $b = 0.060$, $SE = 0.014$, $t = 4.32$, $p < .001$ | **Confidence Interval (significant)** |

*Note*. Within the confidence-interval column, boldface indicates that participants were significantly more likely to follow the advice when it included a confidence interval ($p < .05$). For Study 12, the first row shows the results for the 75%-confidence-interval condition and the second row shows the results for the 95%-confidence-interval condition. See Table 2, for all details on methodological differences between the studies. NBA = National Basketball Association; MLB = Major League Baseball; NFL = National Football League; ns = not significant.

information in the "training block" as a guidance for how many hits are typically scored. In contrast, in Studies 5 and 6, we obtained a significant main effect of this "training" condition in the opposite direction, such that participants were less likely to follow the advice when they saw how the model/advisor had performed previously ($b = -0.051$, $SE = 0.016$, $t = -3.26$, $p = .001$; and $b = -0.041$, $SE = 0.015$, $t = -2.69$, $p = .007$; Dietvorst et al., 2015). Importantly, the interaction between the confidence-interval condition and the training condition was not significant in any of these studies ($ps \geq .477$), indicating that the effect of the confidence-interval condition was not moderated by whether or not information about the model's/advisor's past performance was provided.[4]

### Analysis of Advice Quality

In Study 7, we manipulated the quality of advice that we presented to participants. There was a significant main effect of the advice-quality condition, such that participants were more likely to follow the advice when it was good than when it was random ($b = 0.081$, $SE = 0.015$, $t = 5.43$, $p < .001$). Thus, participants were sensibly sensitive to the quality of the advice. Importantly, however, the interaction between the confidence-interval condition and the advice-quality condition was not significant ($p = .965$).

## Discussion

Studies 1–10 found that participants were either directionally or significantly more likely to choose a model's/advisor's best guess over their own when they saw the best guess accompanied by a confidence interval rather than when they saw only the best guess. Taken together, these studies demonstrate that uncertain advice that includes a confidence interval can indeed be more persuasive than certain advice.

Studies 1–10 were all conducted in the domain of sports. In Study 11, we extended our investigation to a different prediction domain, namely predictions about the death toll of COVID-19.

# Study 11

Participants in Study 11 were asked to predict the number of confirmed deaths due to COVID-19 in the United States by a specified future date. We conducted this study during the COVID-19 pandemic in July 2020.

## Method

### Participants

We conducted Study 11 using U.S. participants from MTurk. Participants received $1 for completing the study, and they could earn an additional $0.25 if their prediction was within 5% of the correct answer. We decided in advance to recruit 4,000 participants for this study. The final data set included all participants who made a final prediction but excluded those whom we preregistered to exclude. As preregistered, we excluded 733 participants with duplicate IP addresses or Turk Prime IDs (regardless of whether or not they passed the attention check that was embedded at the beginning of the study) and 195 participants who were not allowed to continue with the study because they failed the attention check at the beginning of the study. As preregistered, we also excluded 107 participants whose initial or final prediction was lower than the number of confirmed deaths due to COVID-19 in the United States as of the day the study was run (133,290), a number we told participants during the survey. This left us with a final sample of 3,622 participants (average age = 39.0 years; 49.4% female; 49.8% male; 0.8% preferring not to disclose).

### Design

In this study, we again manipulated between subjects whether or not the model's best guess was accompanied by a confidence interval.

### Procedure

Participants were asked to predict the number of confirmed deaths due to COVID-19 in the United States as reported at 7 p.m. Eastern Time on August 1, 2020. We posted this study on July 10, 2020, and we provided participants with the number of confirmed deaths due to COVID-19 in the United States as of July 9, 2020 (133,290; according to Johns Hopkins University and obtained from the website Fivethirtyeight.com). The true answer for this prediction question turned out to be 152,870.

Before making their prediction, participants first learned that they would be asked to provide their prediction rounded to the nearest thousands and without the last three digits (000). That is, we told them to enter their prediction in a number entry box that displayed ",000" to the right of it. We explained this to participants, and then we gave them the number 133,290 (the number of deaths due to COVID-19 as of July 9, 2020) as an example. We asked them to enter the number 133,290 rounded to the nearest thousands in the number entry box that displayed ",000" to the right of it. We specifically told them, "If you wanted to enter the number '133,290' rounded to the nearest thousands, then you would enter the number '133' below." If participants failed to enter the number correctly, they were asked to read the instructions a second time and to try again. If participants failed to enter the number correctly a second time, they were asked to read the instructions a third time and to try again. If participants still entered the number incorrectly after three attempts, they were allowed to continue with the survey. Our final sample included 38 participants who entered the number incorrectly after three attempts.

**Confidence-Interval Manipulation.** Participants were asked to make an initial prediction in the format described above. Then they learned that they would next see the prediction that a statistical model made and that they would have the chance to change their own prediction after seeing the statistical model's advice. We manipulated whether or not the model's best guess was accompanied by a confidence interval in the same way as we did in Studies 1–10. In the no-confidence-interval condition, the model's prediction was presented as a best guess only (e.g., "The statistical model's best prediction is that the number of confirmed deaths due to COVID-19 in the United States will be 146,000."). In the confidence-interval

[4] In Study 5, we did not pre-register to exclude participants who finished the survey after the start of the first game (4:05 p.m.; $n = 132$). Excluding these participants did not alter the direction or significance of the results: Participants were directionally but not significantly more likely to follow the advice when it was accompanied by a confidence interval ($b = 0.013$, $SE = 0.016$, $t = 0.77$, $p = .440$). And they were less likely to follow the advice when they saw how the model had performed previously ($b = -0.054$, $SE = 0.016$, $t = -3.27$, $p = .001$). The interaction was not significant ($p = .342$).

condition, the model's prediction was presented as a best guess with a confidence interval (e.g., "According to the statistical model, there is a 95% chance that the number of confirmed deaths due to COVID-19 in the United States will be between 140,000 and 158,000. Its best prediction is that it will be 146,000.").

We populated the model's prediction in this study with model predictions from the website Fivethirtyeight.com. On the morning of the day the study was run (July 10, 2020), the website offered predictions from 16 different models, all providing a 95% confidence interval around a best prediction. The models' best predictions of what the number of deaths would be on August 1, 2020 ranged from 141,000 (Model 1: Los Alamos) to 166,000 (Model 16: Columbia University; see the online supplemental material 4 for a full list of all model predictions). For stimulus sampling purposes (Wells & Windschitl, 1999), we randomly assigned participants to see the prediction from one of these 16 models. We account for this in our analyses by including fixed effects for model.

**Measuring Advice Following.** After seeing the model's advice, participants were asked to make a final prediction. We told participants, "You now have a chance to change your prediction. Please make your final prediction" (see Figure 2). Participants made their final prediction by typing a number in a textbox rounded to the nearest thousands and without the last three digits (,000), just like they did for their initial prediction. We measured the extent to which participants adjusted their initial prediction after seeing the model's prediction using the same weight-of-advice measure that we used in Study 10.

**Exploratory Measures.** After making their final prediction, participants were asked a series of exploratory questions. The first three measures were similar to the exploratory measures included in Studies 2–10. We asked participants how confident they were in the model's best prediction, how confident they were in their own initial prediction, and how hard they tried to make accurate predictions in this study.[5] In addition, for exploratory purposes, we asked participants, "If you had to vote for Biden or Trump in the elections this fall, who would you vote for?" (5-point scale; 1 = *I would definitely vote for Biden*; 3 = *I don't know*; 5 = *I would definitely vote for Trump*). Finally, we also asked them about their position on wearing masks, "Do you think people should be required to wear masks when they go outside?" (Yes/No/I don't know). Whether people are conservative versus liberal may influence the uptake of the advice, and so we wanted to test whether this moderated our effect.

### *Demographics*

At the end of the study, participants were asked to indicate their age in an open-ended textbox and their gender by selecting from the options "Male", "Female", and "Prefer not to disclose."

## Results

### *Analysis Plan*

As preregistered, we excluded observations for which a participant's initial prediction was identical to the model's best guess (71 participants, 2.0% of the sample) or within 1,000 deaths of the model's best guess (181 participants, 5.1% of the sample). We then winsorized the WOA measure at −1 and 1, as in Study 10 (this affected 228 of the remaining observations, 6.8% of the remaining sample). The final analysis included data from 3,370 participants. We regressed the WOA measure on the confidence-interval condition (1 = confidence interval; 0 = no confidence interval), including fixed effects for model.

### *Main Analysis*

We display the results of Study 11 in Table 3. Participants were significantly more likely to follow the advice when it was accompanied by a confidence interval than when it was not ($WOA_{CI} = 0.52$ vs. $WOA_{NoCI} = 0.48$; $b = 0.038$, $SE = 0.014$, $t = 2.78$, $p = .006$).[6,7,8]

## Discussion

In sum, the results from Study 11 suggest that providing participants with a confidence interval around a best guess can be more persuasive than not providing the confidence interval, not just in the domain of sports, but for other quantitative predictions as well.

## Study 12

In Study 12, we asked participants to predict the preferences of other survey responders, and we presented the advice as coming from a human advisor rather than from a statistical model. Beyond extending our investigation to yet another prediction domain, this

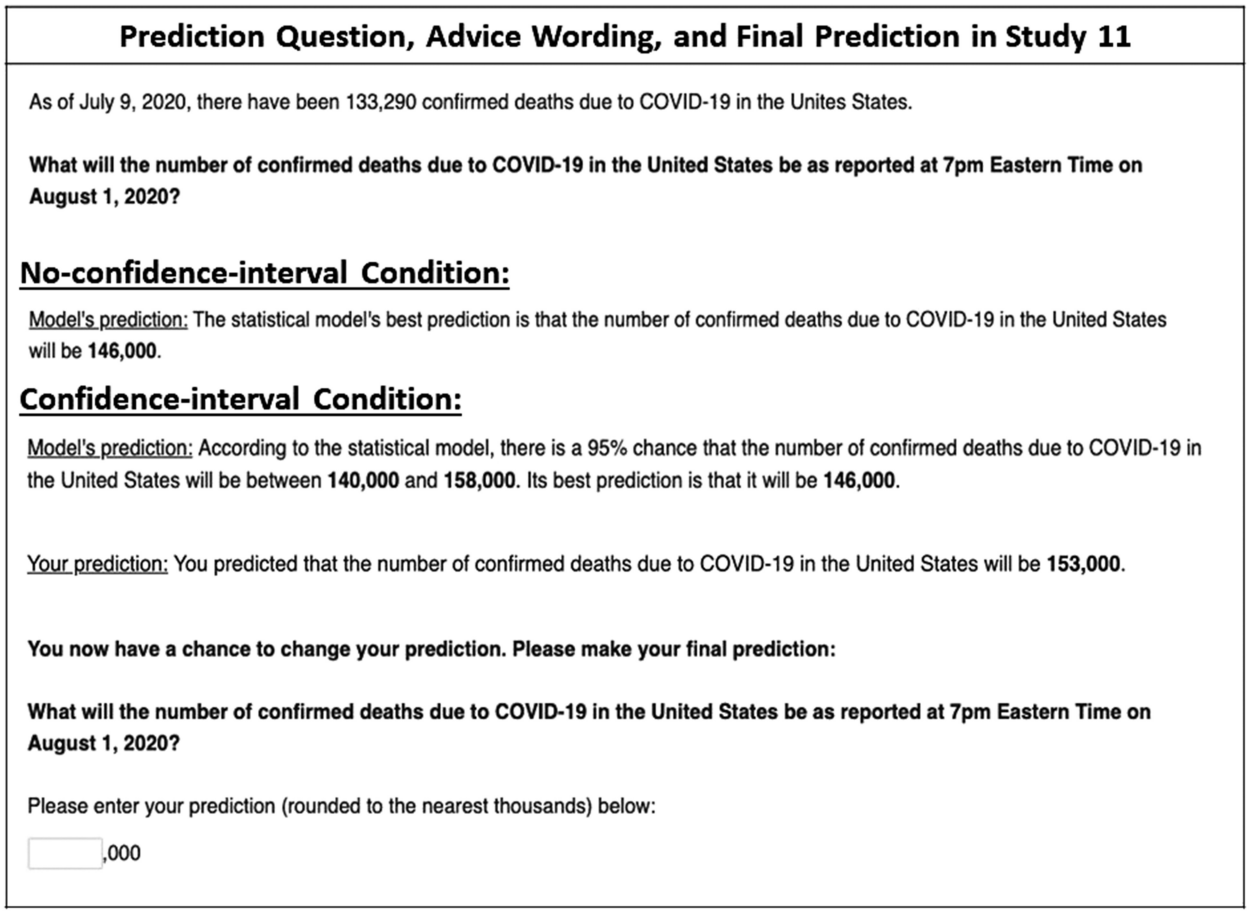
[5] The question that asked participants how hard they had tried to make accurate predictions contained a typo, as it accidently still referred to predicting game outcomes (see Studies 1–10), "How hard did you try to make accurate predictions when making your own predictions about the game outcomes?"

[6] Over the course of conducting our analysis, we noticed that despite the comprehension check that we embedded at the beginning of the study a few participants seemed to not have followed the instructions of entering their prediction rounded to the nearest thousands, as their predictions were very high (e.g., 1.1% of participants gave predictions greater than or equal to one million deaths). To check whether our results hold when excluding these outliers, we conducted additional analyses using different criteria to exclude participants who made extreme predictions. Our effect remains qualitatively unchanged in any of the additional analyses we conducted ($ps \leq .005$). First, we re-ran our main analysis excluding any participants who made an initial or final prediction greater than or equal to one million (1.4% of participants; $b = 0.038$, $SE = 0.013$, $t = 2.86$, $p = .004$). Second, we also re-ran our main analysis excluding any participants who made an initial or final prediction that was more than 50% higher than the high bound of the model to which that participant was assigned (7.4% of participants; $b = 0.037$, $SE = 0.013$, $t = 2.81$, $p = .005$). Third, in an attempt to correct for wrong use of format, we re-ran our main analysis by replacing any prediction greater than or equal to 1,000,000 by that prediction divided by 1,000 (1.1% of each initial predictions and 0.9% of final predictions; $b = 0.043$, $SE = 0.013$, $t = 3.18$, $p = .001$).

[7] As in Studies 1–10, the models' best guesses significantly outperformed participants' predictions, and this was true regardless of how we dealt with outliers. Thus, it was advantageous for participants to follow the model's advice.

[8] We also pre-registered to conduct an ancillary analysis examining whether participants' political affiliation moderated the effect of the confidence-interval condition. We regressed the WOA measure on the advice certainty condition (+0.5 = confidence interval, -0.5 = no confidence interval), the political affiliation measure (mean-centered), and their interaction, including fixed effects for model. Participants were significantly more likely to follow the advice when it was accompanied by a confidence interval than when it was not ($b = 0.038$, $SE = 0.013$, $t = 2.83$, $p = .005$), and when they leaned towards voting for Biden rather than Trump ($b = 0.025$, $SE = 0.004$, $t = 5.85$, $p < .001$; (Joslyn & LeClerc, 2016). The interaction between the confidence interval condition and participants' political affiliation was not significant ($p = .553$).

**Figure 2**
*Prediction Question and Advice Wording in Study 11*




Participants were randomly assigned to only see the model's best guess (no-confidence-interval condition) or to see the best guess accompanied by a confidence interval (confidence-interval condition). The model's advice was populated with advice from models provided on Fivethirtyeight.com, with each participant being randomly assigned to one of 16 different models.



study had two aims. First, we manipulated, between subjects, whether the advice was accompanied by a 75% confidence interval, a 95% confidence interval, or no confidence interval at all. This allowed us to examine whether our prior results are robust to different levels of confidence interval width. Second, before participants made their decision about whether to follow the advice, we asked them two exploratory questions that allowed us to examine whether people are more likely to follow advice accompanied by confidence intervals because of what it signals about the advisor's prediction and/or because of what it signals about their own prediction.

## Method

### *Participants*

We conducted Study 12 on Prolific. Participants received $1.75 for participation. In addition, participants could earn up to an additional $1.50 for accurate prediction performance. Specifically, participants received $0.15, $0.10, or $0.05 for each prediction that was within 1, 3, or 5 of the correct answer, respectively. We decided in advance to recruit 2,000 participants for this study. The final data set included all participants who made a final prediction but excluded those whom we preregistered to exclude. As preregistered, we excluded 49 participants whose IP address or Prolific ID responded to the survey more than once. This left us with a final sample of 1,959 participants (average age = 37.7 years; 48.6% female; 49.6% male; 1.7% nonbinary).

### *Design*

Participants in this study were randomly assigned to one of three between-subject conditions that differed with respect to the format in which the advisor's prediction was presented: No Confidence Interval versus 75% Confidence Interval versus 95% Confidence Interval. We explain the manipulations in detail below.

### *Procedure*

Participants in this study estimated the preferences and behaviors of 100 people we surveyed previously. At the beginning of the survey, participants first answered 10 questions about their own preferences and behaviors, each of which involved a binary choice (e.g., "What

are you more interested in: politics or sports?"). We asked participants to provide their own preferences and behaviors to familiarize them with the items for which we elicited predictions. The 10 questions were presented to participants in a random order on the same screen.

Next, participants were asked to predict how many of 100 people we surveyed previously had given a specific answer to each of those same 10 questions. For example, participants were asked, "How many of the 100 people we surveyed said that they prefer politics to sports?" Table 4 displays all 10 questions that participants made predictions about. We truthfully told participants that we had previously surveyed 100 people, and Table 4 displays the results of that survey, and thus the true answers. Each of the 10 questions was presented one at a time and in a random order.

After making all 10 predictions, participants were told that in the next part of the survey they would see the predictions that an advisor made for each of the questions. We told participants that this advisor previously made predictions for the same questions as they did, but we did not provide them with any more information about the advisor.

We informed participants that they would be asked to indicate which prediction they would like to submit as their official prediction, the advisor's prediction or their own prediction. On the following screens, we presented the 10 questions again, one at a time and in a random order. For each question, we presented participants with the advisor's prediction and we also reminded them of their own prediction. We manipulated the format in which the advisor's prediction was presented.

**Confidence-Interval Manipulation.** Participants were randomly assigned to one of three between-subject conditions: No Confidence Interval versus 75% Confidence Interval versus 95% Confidence Interval. In the no-confidence-interval condition, the advisor's prediction was presented as a best guess only (e.g., "The advisor's best guess was that 56 out of the 100 people surveyed said that they prefer politics to sports."). In the 75%-confidence-interval condition, we added a 75% confidence interval to the best guess (e.g., "The advisor's best guess was that 56 out of the 100 people surveyed said that they prefer politics to sports. The advisor also said there is a 75% chance that the true answer is between 50 and 62."). And in the 95%-confidence-interval condition, we added a 95% confidence interval to the best guess (e.g., "The advisor's best guess was that 56 out of the 100 people surveyed said that they prefer politics to sports. The advisor also said there is a 95% chance that the true answer is between 46 and 66."). The wording of these conditions is also displayed in Figure 3.

For each of the 10 preference questions, we generated the advisor's best guess based on the true answers from a pilot study. In order to make our stimuli more realistic, for each item, we randomly assigned participants to one of five types of advisor best guess: exact true answer, true answer minus 5, true answer plus 5, true answer minus 11, and true answer plus 11. Thus, for a given item, the advisor's best guess was either 0, 5, or 11 away from the true answer. For example, the true response for the number of previous participants who prefer politics to sports was 56 out of 100, and so we randomly assigned participants to a best guess of 56, 51, 61, 45, or 67.

We generated the confidence intervals around the best guesses using margin of error calculations. Across items and depending on the exact true answer, the 75% confidence interval encompassed a distance of either 5 or 6 from the best guess, and the 95% confidence interval encompassed a distance of between 7 and 10 from the best guess. The online supplemental material 5 provides the best guesses for all items along with the corresponding 75% and 95% confidence intervals.

**Table 4**
*Prediction Questions and True Answers in Study 12*

| Number of participants who … | True answer (out of 100) |
|---|---|
| … prefer peanut butter cookies to chocolate chip cookies | 26 |
| … drink at least 10 cups of coffee in a week | 30 |
| … prefer having a cat to having a dog | 39 |
| … own an iPad | 41 |
| … prefer vanilla ice cream to chocolate ice cream | 44 |
| … have posted a video on YouTube | 53 |
| … prefer Spring to Summer | 56 |
| … prefer politics to sports | 56 |
| … have traveled outside the United States | 71 |
| … have a Twitter account | 74 |

**Mediator Measures.** We displayed participants' own prediction immediately below the advisor's advice. We then asked participants to answer two questions about the advisor's prediction and their own prediction. Specifically, for each prediction question, we asked participants to indicate how accurate they thought the advisor's prediction was for this question (1 = Very inaccurate; 9 = Very accurate), and how accurate they thought their own prediction was for this question (1 = Very inaccurate; 9 = Very accurate). We included these questions as potential mediators in this study.

### *Advice Following*

Below these questions and on the same page, we asked participants to indicate which prediction they would like to submit as their official prediction. For example, for the item presented in Figure 3, we asked, "Which prediction would you like to submit as your official prediction–the advisor's (56) or your own (48)?" Participants indicated their choice by selecting one of the two options. We measured advice following by calculating the percentage of observations for which participants chose the advisor's best guess over their own prediction.

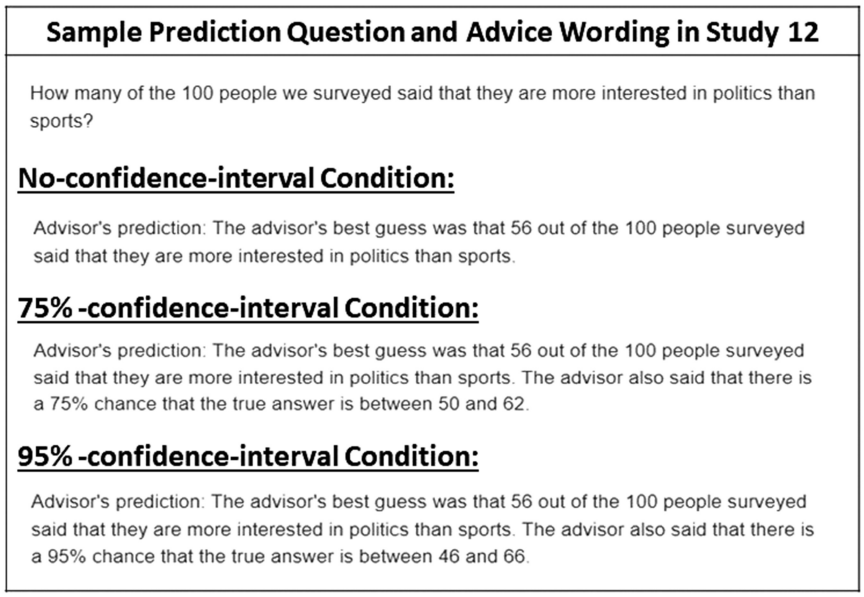
**Exploratory Measures and Demographics.** At the end of the survey, we asked participants to indicate how hard they tried when making their predictions (1 = I did not try hard at all; 7 = I tried extremely hard). Finally, participants were asked to indicate their age by selecting an answer from a dropdown list and their gender by selecting from the options "Male", "Female", and "Nonbinary."

## Results

### *Analysis Plan*

Each participant contributed 10 observations to the dataset, one for each of the preferences/behaviors they predicted. As preregistered, we dropped those observations for which a participant's prediction was identical to the advisor's best guess ($n = 307$; 1.6% of observations). First, we regressed participants' choice of the advisor's best guess (1 = they chose the advisor's best guess, 0 = they did not choose the advisor's best guess) on (1) the 75%-confidence-interval condition and (2) the 95%-confidence-interval condition. This regression allowed us to compare each of the two confidence-interval conditions

**Figure 3**
*Advice Wording for one of the Items in Study 12 Across the Three Conditions*




Participants were randomly assigned to only see the advisor's best guess (no-confidence-interval condition) or to see the best guess accompanied by either a 75% or a 95% confidence interval (75%-confidence-interval condition and 95%-confidence-interval condition). In addition, for each item, we randomly assigned participants to one of five types of advisor best guess: exact true answer, true answer minus 5, true answer plus 5, true answer minus 11, and true answer plus 11. That is, the advisor's best guess was either 0, 5, or 11 away from the true answer. In the screenshot above, the advisor's best guess reflects the exact true answer for this question.



to the no-confidence-interval condition. Second, we regressed participants' choice of the advisor's best guess (1 = they chose the advisor's best guess, 0 = they did not choose the advisor's best guess) on (1) the 75%-confidence-interval condition, and (2) the no-confidence-interval condition. This regression allowed us to compare the two confidence-interval conditions to each other. In each of these regressions, we included fixed effects for item and for the distance of the advisor's prediction from the true answer (0, 5, or 11), and we clustered standard errors by participant.

### *Main Analysis*

Participants were significantly more likely to choose the advisor's prediction over their own when it was accompanied by a 75% confidence interval ($M = 40.6\%$; $b = 0.058$, $SE = 0.014$, $t = 4.23$, $p < .001$) and when it was accompanied by a 95% confidence interval ($M = 40.8\%$; $b = 0.060$, $SE = 0.014$, $t = 4.32$, $p < .001$) than when no confidence interval was presented ($M = 34.8\%$; see the bottom row of Table 3). The two confidence-interval conditions did not significantly differ from each other ($b = 0.002$, $SE = 0.014$, $t = 0.14$, $p = .891$).

**Perceived Accuracy of the Advisor's and One's Own Predictions.** Despite the fact that the advisor provided identical best predictions across all three conditions, exploratory analyses revealed that participants rated the advisor's predictions as significantly more accurate when either a 75% confidence interval ($M = 5.67$, $SD = 1.85$; $b = 0.298$, $SE = 0.058$, $t = 5.11$, $p < .001$) or a 95% confidence interval ($M = 5.77$, $SD = 1.92$; $b = 0.407$, $SE = 0.060$, $t = 6.79$, $p < .001$) was presented around the advisor's best prediction compared to when no confidence interval was presented ($M = 5.37$, $SD = 1.93$). The two confidence-interval conditions did not significantly differ from each other ($b = 0.109$, $SE = 0.060$, $t = 1.83$, $p = .068$), though there was a slight tendency to judge the advice accompanied by a (wider) 95% confidence interval as more accurate than the advice accompanied by a (narrower) 75% confidence interval.

Participants' perceptions of the accuracy of their own predictions were unaffected by whether or not the advisor's best prediction was accompanied by a confidence interval ($M_{\text{NoCI}} = 5.74$, $SD = 1.61$ vs. $M_{75\%\text{CI}} = 5.69$, $SD = 1.67$ vs. $M_{95\%\text{CI}} = 5.71$, $SD = 1.79$; all $ps \geq .361$). Thus, the presence of a confidence interval altered participants' perceptions of the advisor's performance, but not of their own performance.

### *Mediation Analysis*

We next tested whether participants' perceptions of the advisor's performance mediated the effect of the confidence interval on participants' choice. For this mediation analysis, which was not preregistered, we collapsed the two confidence interval conditions into one condition. We found that the effect of confidence-interval condition on whether participants chose to follow the advisor's advice was significantly lower when participants' judgments about the advisor's performance were included in the model ($b = 0.021$, $SE = 0.010$,

$t = 2.04$, $p = .042$) compared to when they were not ($b = 0.059$, $SE = 0.012$, $t = 4.95$, $p < .001$, 95% bootstrapped confidence interval for the difference = [0.028, 0.051]).[9]

Thus, these exploratory and correlational analyses suggest that participants may have been more likely to follow advice accompanied by confidence intervals in part because they believed that advice to have been more accurate.

## General Discussion

In 11 out of 12 studies, we found that people were either directionally or significantly more likely to follow numeric advice when this advice was accompanied by a confidence interval around a best guess. All studies were incentivized, and the results were robust to different measures of advice following, including asking participants to make one-at-a-time choices between their own prediction and an advisor's best guess, asking them to make a choice once for all the predictions they made, or measuring the extent to which they adjusted their own prediction in the direction of an advisor's best guess. Our results seem not to depend on whether the width of the confidence interval is 75% or 95% (Study 12), whether the quality of the advice is good or random (Study 7), or on whether people have information on how the model previously performed (Studies 4–6). Taken together, our results suggest that numerical advice may be more persuasive when it is accompanied by a confidence interval.

Of course, these findings raise the question of *why* people are more likely to follow advice that is accompanied by a confidence interval. Our data do not allow us to examine all possible mechanisms, but we were able to conduct some exploratory analyses that examine a few possibilities.

First, of the exploratory variables that we analyzed, confidence in the model/advisor was the only one that was somewhat reliably influenced by our manipulation (see the online supplemental material 2). Specifically, we often found that providing a confidence interval increased confidence in the model/advisor. On the one hand, this effect was significant in only five of the eleven studies. Moreover, because it was assessed after the advice-taking measures, it could be a consequence rather than a cause of this effect. On the other hand, in Study 12, we found that participants believed that advice was more accurate when it was accompanied by a confidence interval, and that these beliefs mediated the effect of confidence-interval presentation on advice following. In total, the balance of evidence suggests that people are more likely to follow advice accompanied by confidence intervals in part because they are more confident in or trusting of that advice.

Why would people be more trusting of advice accompanied by a confidence interval? In all of our studies, participants were asked to forecast events that were inherently uncertain, including future sporting events, the future trajectory of the COVID-19 pandemic, and the preferences of other survey responders. When receiving advice about uncertain events, advisees may recognize that the events are indeed uncertain and may deem forecasts that incorporate that uncertainty to be more credible (Budescu & Wallsten, 1995; Du et al., 2011). Consistent with this, prior research has revealed that providing uncertainty in the form of a range or confidence interval can lead people to perceive an advisor's forecasts to be more accurate, credible, and trustworthy (e.g., Du et al., 2011; Gaertig & Simmons, 2018; Joslyn & LeClerc, 2012, 2016).

Indeed, the fact that in Study 12 the width of the confidence interval (75% or 95%) did not influence participants' uptake of the uncertain advice suggests that perhaps what people value most about (reasonably sized) confidence intervals is the signal that the advisor is intelligently incorporating uncertainty; how much of that uncertainty the advisor communicates may be less important, at least when the confidence intervals are not so narrow or wide as to be uninformative.

But the effect of confidence intervals on people's perceptions of the advisor's/advice's credibility may not be the only mechanism at play here. Another possibility is that confidence intervals may increase advice following by allowing people to recognize that their own judgment is so far from the truth that it should be revised. For example, those who see that their judgment is outside of a 75% or 95% confidence interval may be more likely to recognize that their judgment is likely to be errant, thereby making them more receptive to an advisor's advice.

In Studies 1–12, participants first made their own prediction and then received advice. Prior research has found that the extent to which people take advice depends on the distance between their own estimate and the advice (Schultze et al., 2015). In our studies, we not only showed participants the advisor's best guess, but we also manipulated the presence or absence of a confidence interval around the advice. Participants in the confidence-interval condition were therefore able to observe whether their own prediction was within the confidence interval. If people are more likely to follow advice that is accompanied by a confidence interval precisely because this form of advice allows them to better appreciate when their own judgment is likely to be very off the mark, then our effect should be strongest among those whose initial judgments were outside of the confidence interval.

Table 5 shows the results of Studies 1, 2, and 4–12 split up by whether participants' initial prediction was within or outside of the confidence interval. These exploratory analyses include all studies except for Study 3, for which this analysis was impossible because participants made one choice for all games. Note that in the no-confidence-interval condition, participants never saw the advisor's confidence interval, and so our data split is based on the confidence interval that participants in the confidence-interval condition saw. Study 12 included two different confidence-interval conditions (75% and 95%), and we conducted separate analyses for each of those.

As can be seen in Table 5, in all but two studies (Studies 11 and 12) the majority of participants' initial predictions were within the confidence interval (between 70.4% and 94.6% of observations in Studies 1–10 versus between 19.5% and 30.9% of observations in Studies 11 and 12).

The top half of Table 5 shows that when participants' initial predictions were within the confidence interval, participants in seven of the eleven studies were still directionally more likely to follow the advice that was accompanied by a confidence interval, and this effect was significant in two of these seven studies. Interestingly, however, the effect of the confidence-interval condition on advice following seems to be more

[9] These analyses clustered standard errors by participant, and included fixed effects for item and for the absolute difference between the advice and the truth. The 95% bootstrapped confidence interval was based on 1,000 re-samples.

pronounced for participants whose initial predictions were outside of the confidence interval, with eight of the eleven studies yielding a significant effect of the confidence-interval condition (see Table 5 bottom half). This suggests that providing a confidence interval may be more effective for those whose initial predictions are far enough from the advisor's best guess so as to be outside of the confidence interval. These, of course, are the people who would most benefit from following good advice.

## Limitations and Future Directions

Our research leaves many interesting and important questions unanswered. First, in all of our studies, we focus on presenting confidence intervals in numeric terms. Of course, one can express uncertainty in many different ways. For example, uncertainty can also be expressed in verbal terms (e.g., "this outcome is likely") and people often use verbal probabilities differently than numeric probabilities

**Table 5**

*Results for Studies 1, 2, and 4–12 for Observations for Which Participants' Initial Predictions Were Inside or Outside of the Confidence Interval*

| Participants' initial predictions were inside the confidence interval | | | | | | | |
|---|---|---|---|---|---|---|---|
| Study | Final study $n$ | Percent of observations | Domain | Measure of advice following | No confidence interval | Confidence interval | Main effect of confidence interval (vs. no confidence interval) |
| 1 | 381 | 92.4% | NBA | Percent of time participant chose the model's best guess over their own prediction | 30.8% | 34.3% | $b = 0.035$, $SE = 0.031$, $t = 1.10$, $p = .270$ |
| 2 | 377 | 94.6% | | | 18.6% | **24.1%** | $b = 0.056$, $SE = 0.026$, $t = 2.17$, $p = .030$ |
| 4 | 1,076 | 80.2% | MLB | Percent of time participant chose the model's/advisor's best guess over their own prediction | 43.2% | 41.8% | $b = -0.016$, $SE = 0.019$, $t = -0.84$, $p = .401$ |
| 5 | 1,222 | 93.1% | | | 37.6% | 37.2% | $b = -0.004$, $SE = 0.016$, $t = -0.26$, $p = .796$ |
| 6 | 1,233 | 75.4% | | | 26.2% | 28.0% | $b = 0.018$, $SE = 0.014$, $t = 1.25$, $p = .210$ |
| 7 | 1,205 | 81.9% | | | 28.1% | 28.8% | $b = 0.005$, $SE = 0.015$, $t = 0.36$, $p = .722$ |
| 8 | 1,156 | 70.4% | NFL | Percent of time participant chose the model's best guess over their own prediction | 26.0% | 27.5% | $b = 0.015$, $SE = 0.018$, $t = 0.84$, $p = .402$ |
| 9 | 3,842 | 79.9% | | | 23.6% | **26.5%** | $b = 0.029$, $SE = 0.009$, $t = 3.20$, $p = .001$ |
| 10 | 1,166 | 80.7% | | Weight of advice for each game | 0.24 | 0.26 | $b = 0.021$, $SE = 0.015$, $t = 1.37$, $p = .172$ |
| 11 | 3,622 | 23.6% | COVID-19 | Weight of advice | 0.43 | *0.37* | $b = -0.062$, $SE = 0.026$, $t = -2.33$, $p = .020$ |
| 12 | 1,959 | 19.5% | Preferences | Percent of time participant chose the advisor's best guess over their own prediction | 25.4% | 23.4% | $b = -0.019$, $SE = 0.021$, $t = -0.93$, $p = .352$ |
| | | 30.9% | | | 31.7% | 28.2% | $b = -0.035$, $SE = 0.019$, $t = -1.82$, $p = .069$ |
| Participants' initial predictions were outside of the confidence interval | | | | | | | |
| Study | Final study $n$ | Percent of observations | Domain | Measure of advice following | No confidence interval | Confidence interval | Main effect of confidence interval (vs. no confidence interval) |
| 1 | 381 | 7.6% | NBA | Percent of time participant chose the model's best guess over their own prediction | 49.6% | 48.8% | $b = -0.002$, $SE = 0.089$, $t = -0.02$, $p = .984$ |
| 2 | 377 | 5.4% | | | 20.8% | **41.9%** | $b = 0.213$, $SE = 0.074$, $t = 2.88$, $p = .005$ |
| 4 | 1,076 | 19.8% | MLB | Percent of time participant chose the model's/advisor's best guess over their own prediction | 79.1% | 82.9% | $b = 0.025$, $SE = 0.026$, $t = 0.94$, $p = .345$ |
| 5 | 1,222 | 6.9% | | | 69.8% | **81.2%** | $b = 0.118$, $SE = 0.038$, $t = 3.13$, $p = .002$ |
| 6 | 1,233 | 24.6% | | | 45.8% | **58.8%** | $b = 0.129$, $SE = 0.026$, $t = 4.96$, $p < .001$ |
| 7 | 1,205 | 18.1% | | | 39.4% | **48.2%** | $b = 0.128$, $SE = 0.034$, $t = 3.76$, $p < .001$ |
| 8 | 1,156 | 29.6% | NFL | Percent of time participant chose the model's best guess over their own prediction | 57.8% | **66.2%** | $b = 0.084$, $SE = 0.027$, $t = 3.06$, $p = .002$ |
| 9 | 3,842 | 20.1% | | | 56.2% | **60.5%** | $b = 0.043$, $SE = 0.018$, $t = 2.35$, $p = .019$ |
| 10 | 1,166 | 19.3% | | Weight of advice for each game | 0.53 | 0.57 | $b = 0.050$, $SE = 0.031$, $t = 1.57$, $p = .116$ |
| 11 | 3,622 | 76.4% | COVID-19 | Weight of advice | 0.49 | **0.56** | $b = 0.072$, $SE = 0.015$, $t = 4.66$, $p < .001$ |
| 12 | 1,959 | 80.5% | Preferences | Percent of time participant chose the advisor's best guess over their own prediction | 37.1% | **44.8%** | $b = 0.077$, $SE = 0.015$, $t = 5.13$, $p < .001$ |
| | | 69.1% | | | 36.3% | **46.3%** | $b = 0.100$, $SE = 0.016$, $t = 6.29$, $p < .001$ |

*Note.* Within the confidence-interval column, boldface indicates that participants were significantly more likely to follow the advice when it included a confidence interval ($p < .05$), and italics indicates that participants were significantly more likely to follow the advice when it did not include a confidence interval ($p < .05$). Study 3 is not included in this table, since participants made only one choice for all games in this study. For Study 12, the first row shows the results for the 75%-confidence-interval condition and the second row shows the results for the 95%-confidence-interval condition. NBA = National Basketball Association; MLB = Major League Baseball; NFL = National Football League.

(e.g., Mislavsky & Gaertig, 2022; Windschitl & Weber, 1999; for a recent review see Dhami & Mandel, 2022). Future work could help us more fully understand how different expressions of uncertainty affect people's tendency to follow an advisor's advice.

Second, participants in our studies did not receive any feedback as to whether or not their forecasts were correct. It is unclear whether providing such feedback would increase or decrease participants' trust of advisors who provide calibrated confidence intervals. On the one hand, seeing that the confidence intervals are calibrated may increase trust (Tenney et al., 2007). On the other hand, witnessing the cases in which the true answer lies outside of the confidence interval may decrease trust, especially when confidence intervals are narrower (e.g., 75% rather than 95%) and thus those instances are more common. The fact that we did not observe different results when participants saw how advisors (and their confidence intervals) had previously performed suggests that this may not happen. But the results could be different when participants receive feedback after they themselves have followed or rejected the advice.

Third, although our work, together with other recent research, demonstrates an appreciation of uncertainty across different inherently uncertain domains, such as sports (Gaertig & Simmons, 2018), weather (Joslyn & LeClerc, 2012, 2016), and financial forecasting (Du et al., 2011), many other domains exist in which people face inherently uncertain outcomes. In some of those domains, people may be more averse to uncertain information. For example, when facing medical decisions, people may be less likely to tolerate uncertainty and may prefer experts who add clear recommendations to their forecast (Kassirer et al., 2020).

Finally, all of our research was conducted using online samples of participants based in the United States. Although U.S.-based online samples are much more representative than U.S.-based college samples, they obviously do not capture the psychology of people from vastly different countries or cultures. Indeed, it is possible that people's tolerance of or preferences for uncertainty differ across countries or cultures, and that our results do not generalize to every country or culture. We look forward to future research investigating how these findings generalize to other populations.

## Conclusion

In our research, we find that people are more likely than not to follow numerical advice that is accompanied by a confidence interval. Alongside other recently published findings (e.g., Du et al., 2011; Gaertig & Simmons, 2018; Gustafson & Rice, 2019; Joslyn & LeClerc, 2012, 2016), these results suggest that uncertain information is not inherently distasteful. Indeed, rather than merely tolerating uncertainty, most people may actually prefer to heed the advice of those who accurately communicate it.

## References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). Cambridge University Press.

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Best, R., & Boice, J. (2020, December 4). *Where the latest COVID-19 models think we're headed—And why they disagree*. FiveThirtyEight. https://projects.fivethirtyeight.com/covid-forecasts/

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation*, *32*, 275–318. https://doi.org/10.1016/S0079-7421(08)60313-8

Della Vigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Science*, *26*(6), 514–526. https://doi.org/10.1016/j.tics.2022.03.002

Dietvorst, B., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Du, N., Budescu, D. V., Shelley, M., & Omer, T. C. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes*, *114*(2), 179–189. https://doi.org/10.1016/j.obhdp.2010.10.005

Gaertig, C., & Simmons, J. P. (2018). Do people inherently dislike uncertain advice? *Psychological Science*, *29*(4), 504–520. https://doi.org/10.1177/0956797617739369

Gustafson, A., & Rice, R. E. (2019). The effects of uncertainty frames in three science communication topics. *Science Communication*, *41*(6), 679–706. https://doi.org/10.1177/1075547019870811

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Howe, L. C., MacInnis, B., Krosnick, J. A., Markowitz, E. M., & Socolow, R. (2019). Acknowledging uncertainty impacts public acceptance of climate scientists' predictions. *Nature Climate Change*, *9*(11), 863–867. https://doi.org/10.1038/s41558-019-0587-5

Joslyn, S. L., & LeClerc, J. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126–140. https://doi.org/10.1037/a0025185

Joslyn, S. L., & LeClerc, J. E. (2016). Climate projections and uncertainty communication. *Topics in Cognitive Science*, *8*(1), 222–241. https://doi.org/10.1111/tops.12177

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Strauss, Giroux.

Kassirer, S., Levine, E., & Gaertig, C. (2020). Decisional autonomy undermines advisees' judgments of experts in medicine and in life. *Proceedings of the National Academy of Sciences*, *117*(21), 11368–11378. https://doi.org/10.1073/pnas.1910572117

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247. https://doi.org/10.1006/obhd.1999.2847

Mislavsky, R., & Gaertig, C. (2022). Combining probability forecasts: 60% and 60% is 60%, but likely and likely is very likely. *Management Science*, *68*(1), 541–563. https://doi.org/10.1287/mnsc.2020.3902

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moore, D. A., Tenney, E. R., & Haran, U. (2016). Overprecision in judgment. In G. Keren & G. Wu (Eds.), *Wiley Blackwell handbook of judgment and decision making* (pp. 182–209). Wiley.

National Academies of Sciences, Engineering, and Medicine. (2017). *Communicating science effectively: A research agenda*. The National Academies Press.

Nickerson, W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making.

*Psychological Science*, *21*(2), 194–199. https://doi.org/10.1177/0956797609359326

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

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57. https://doi.org/10.1002/bdm.460

Radzevick, J. R., & Moore, D. A. (2011). Competing to be certain (but wrong): Market dynamics and excessive confidence in judgment. *Management Science*, *57*(1), 93–106. https://doi.org/10.1287/mnsc.1100.1255

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, *121*(2), 246–255. https://doi.org/10.1016/j.obhdp.2013.02.001

Schultze, T., Rakotoarisoa, A., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*(2), 144–171. https://doi.org/10.1017/S1930297500003922

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor-system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. https://doi.org/10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368–1375. https://doi.org/10.1016/j.jesp.2008.04.006

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.

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

Van Zant, A. B. (2022). Strategically overconfident (to a fault): How self-promotion motivates advisor confidence. *Journal of Applied Psychology*, *107*(1), 109–129. https://doi.org/10.1037/apl0000879

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of likely" depends on the context, but "70%" is 70%—right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1514–1533. https://doi.org/10.1037/0278-7393.25.6.1514

Yaniv, I. (2004). 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

Received January 6, 2022
Revision received January 10, 2023
Accepted January 19, 2023 ■