Journal of Experimental Psychology: General

ISSN: 0096-3445

2025, Vol. 154, No. 2, 476–496
https://doi.org/10.1037/xge0001655

Different Methods Elicit Different Belief Distributions

Beidi Hu[1] and Joseph P. Simmons[2]
[1] Marketing Group, The University of Chicago Booth School of Business
[2] Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania

When eliciting people's forecasts or beliefs, you can ask for a point estimate—for example, what is the most likely state of the world?—or you can ask for an entire distribution of beliefs—for example, how likely is *every* possible state of the world? Eliciting belief distributions potentially yields more information, and researchers have increasingly tried to do so. In this article, we show that different elicitation methods elicit different belief distributions. We compare two popular methods used to elicit belief distributions: Distribution Builder and Sliders. In 10 preregistered studies ($N = 14{,}553$), we find that Distribution Builder elicits more accurate belief distributions than Sliders, except when true distributions are right-skewed, for which the results are mixed. This result holds when we assess accuracy (a) relative to a normative benchmark and (b) relative to participants' own beliefs. Our evidence suggests that participants approach these two methods differently: Sliders users are more likely to start with the lowest bins in the interface, which in turn leads them to put excessive mass in those bins. Our research sheds light on the process by which people construct belief distributions while offering a practical recommendation for future research: All else equal, Distribution Builder yields more accurate belief distributions.

Public Significance Statement

The practice of eliciting people's entire belief distributions has been on the rise among researchers in psychology and adjacent fields, and the extant literature implicitly assumes that people generate the same belief distribution regardless of which method they use. In a comparison between two popular methods, we find that Distribution Builder elicits more accurate distributions than Sliders most of the time. This suggests that, as with elicitations of choices, point estimates, and interval estimates, researchers should be mindful of how the elicitation method may alter people's belief distributions and the conclusions drawn. Overall, our research sheds light on the process by which people construct belief distributions and offers a practical recommendation for future research: All else equal, Distribution Builder yields more accurate belief distributions.

Keywords: belief distributions, belief elicitation, response format, judgmental biases

Supplemental materials: https://doi.org/10.1037/xge0001655.supp

How should researchers assess people's beliefs? By far, the most common approach is to ask them to report a single number, such as a point estimate (e.g., "What do you think Company X's stock price is going to be in a year?") or a likelihood judgment (e.g., "How likely is Company X's stock price to increase over the next year?"). But these simple reports omit a great deal of information. Most notably, they fail to capture people's beliefs about all possible alternatives (e.g., "How likely is Company X's stock to go up or down by 0%–2%, 2%–4%, 4%–6%, etc.?"). To remedy this, researchers have recently embraced the practice of eliciting people's *subjective belief distributions*, essentially asking them to report their beliefs about all possible outcomes (André et al., 2022; Bechler & Levav, 2022; Camilleri et al., 2019; Delavande & Rohwedder, 2008; Dietvorst & Bharti, 2020; Dimant et al., 2024; Goldstein & Rothschild, 2014; Haran et al., 2010; Hofman et al., 2020; Hu & Simmons, 2023; Leemann et al., 2021; Lim et al., 2021; Long et al., 2018;

This article was published Online First September 26, 2024.
Abigail Sussman served as action editor.
Beidi Hu https://orcid.org/0000-0002-8190-716X
All of our data, materials, and preregistrations are available on ResearchBox at https://researchbox.org/569. Part of the data in this article has been presented at the 2020 *Society for Judgment and Decision Making* conference, the 2021 *Association for Consumer Research* conference, and the 2022 *Society for Consumer Psychology* conference. This research was conducted as part of Beidi Hu's dissertation. This research was supported by the Wharton Behavioral Lab, the Wharton Dean's Fund, and the Wharton Risk Management Center's Russell Ackoff Doctoral Student Fellowship awarded to Beidi Hu.

Beidi Hu played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, visualization, and writing–original draft and an equal role in funding acquisition and writing–review and editing. Joseph P. Simmons played a lead role in supervision, a supporting role in conceptualization, data curation, formal analysis, investigation, and methodology, and an equal role in funding acquisition and writing–review and editing.

Correspondence concerning this article should be addressed to Beidi Hu, Marketing Group, The University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, United States. Email: beidi.hu@chicagobooth.edu

Mannes & Moore, 2013; Moore et al., 2015, 2017; Muthukrishna et al., 2018; Page & Goldstein, 2016; Prims & Moore, 2017; Reinholtz et al., 2021; Ren & Croson, 2013; Robbins & Hemmer, 2018; Soll et al., 2023; Zhang et al., 2023). In this article, we compare two common ways of eliciting belief distributions and find that one of them is usually better. In so doing, we shed light on the psychology underlying the construction of belief distributions and uncover ways to alter or improve the distributions resulting from one common method of elicitation.[1]

By eliciting belief distributions, researchers may gain access to much more information about people's beliefs. For example, imagine that a researcher asks someone to forecast the outcome of a baseball game played between the Baltimore Orioles and the New York Yankees. A researcher who simply asks who is going to win and by how many runs[2] might learn that the participant believes that the Orioles are going to beat the Yankees by 2 runs. But the researcher who elicits the participant's subjective belief distribution—asking the participant to indicate the likelihood of *all* possible game outcomes—might also learn that the participant believes that (a) "Orioles by 2" is the modal outcome, but that "Orioles by 3" is the *average* outcome; (b) the Orioles have a 75% chance to win the game; (c) there is a 95% chance that the game will be decided by fewer than 5 runs; and so on.

The richness of this information offers the obvious (potential) advantage of allowing researchers to more thoroughly understand multiple aspects of participants' beliefs. And that advantage begets others. For example, belief distribution data can allow researchers to ensure that their results are robust to different derivations of a statistical property (e.g., multiple ways of calculating dispersion, such as range, standard deviation, or absolute deviation; e.g., André et al., 2022; Bechler & Levav, 2022; Reinholtz et al., 2021). Belief distribution data may also provide researchers with greater precision for fitting formal functions in econometric models (Manski, 2004) and Bayesian decision models (Garthwaite et al., 2005). Finally, eliciting belief distributions may sidestep concerns about the validity of single survey questions and, as prior work suggests, more accurately capture participants' beliefs (Goldstein & Rothschild, 2014). For example, questions that require participants to report their beliefs about central tendency or variance often require respondents to convert their subjective distributions into single statistics, a process potentially plagued with biases and noise that may muddle the interpretation of results. For all these reasons, researchers have increasingly asked participants to provide belief distributions rather than point estimates.

To elicit subjective belief distributions, researchers typically provide an interface that invites respondents to produce a histogram of frequencies or probabilities across all possible answers or outcomes (Goldstein & Rothschild, 2014; Haran et al., 2010). Specifically, the interface (a) divides the entire range of possible options into several mutually exclusive and collectively exhaustive bins and (b) asks participants to estimate the frequency of each bin or the probability that the true outcome will fall into each bin.

Two graphical variants of this approach have most frequently been used: the *Distribution Builder* method proposed by Goldstein and Rothschild (2014; see also Goldstein et al., 2008; Sharpe et al., 2000) and the *Subjective Probability Interval Estimates* method proposed by Haran et al. (2010). Both methods allow participants to create visual histograms that represent their subjective distributions. The key difference between the two methods is the interface. The Distribution Builder (Goldstein & Rothschild, 2014) uses a graphical interface that allows participants to allocate a fixed number of balls into vertically displayed bins, each representing one bin of the range (Figure 1 Panel A presents an example using the *distBuilder* tool developed by André, 2016). The Subjective Probability Interval Estimates interface (Haran et al., 2010; hereafter referred to as "Sliders") uses an ordered array of horizontal slider scales, each representing one bin (Figure 1 Panel B presents an example). Participants assign probabilities or frequencies to each bin by sliding the bars from left to right.

Both methods have been used often. The Distribution Builder was first introduced as a tool for marketers and investment managers to elicit customers' desired outcome distributions (Goldstein et al., 2008; Sharpe et al., 2000). In addition to receiving continued attention from researchers examining risk preferences in the domains of retirement savings (Bokern et al., 2021; Camilleri et al., 2019; Donkers et al., 2013) and consumer investments (Lim et al., 2021), Distribution Builder and similar ball-and-bin methods have been used to obtain fine-grained information about people's statistical intuitions (André et al., 2022; Bechler & Levav, 2022; Goldstein & Rothschild, 2014; Hofman et al., 2020; Zhang et al., 2023), as well as to explore the psychology underlying people's financial decisions (Long et al., 2018; Reinholtz et al., 2021), choice of forecasting methods (Dietvorst & Bharti, 2020), and norm perception (Dimant, 2023; Dimant et al., 2024). Further, it has garnered attention in other disciplines to study people's beliefs about social security benefits (Delavande & Rohwedder, 2008), income distributions (Page & Goldstein, 2016), and voter expectations (Leemann et al., 2021). The Sliders interface, on the other hand, has been primarily used in research investigating overconfidence (Camilleri & Newell, 2019; Haran et al., 2010; Hu & Simmons, 2023; Mannes & Moore, 2013; Moore et al., 2015; Muthukrishna et al., 2018; Prims & Moore, 2017; Ren & Croson, 2013; Soll et al., 2023) and has been adopted in geopolitical forecasting (Moore et al., 2017).

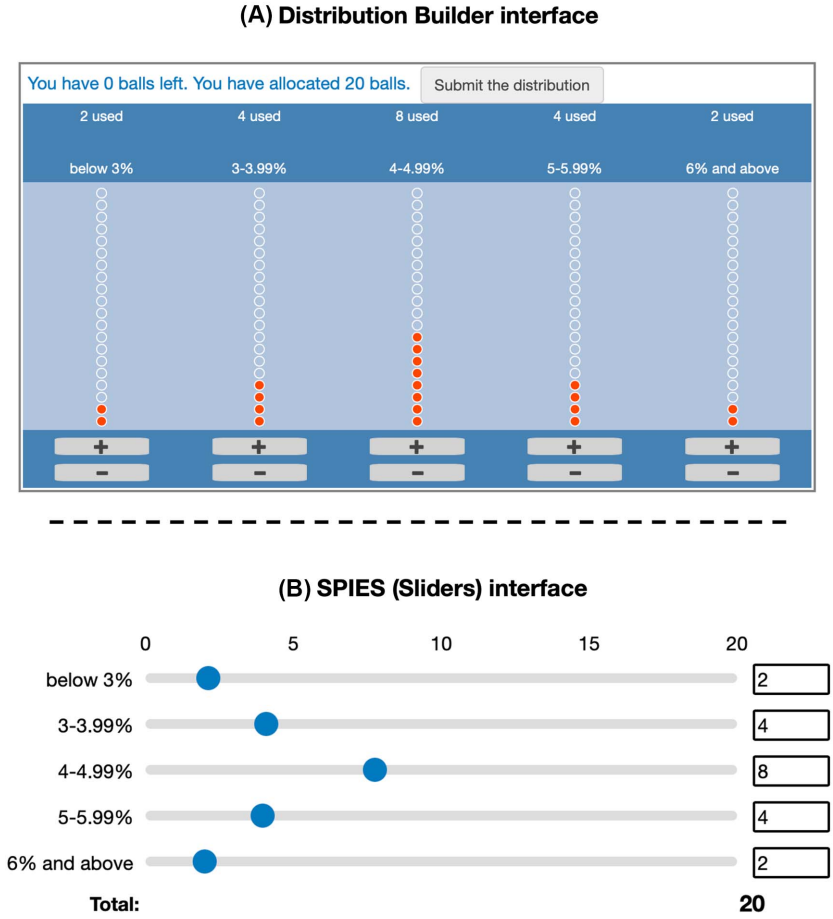
Beyond these graphical methods, eliciting entire belief distributions has a long history in professional forecasting and nationwide surveys, both of which help inform economic and financial policy decisions. For instance, since 1968, the Survey of Professional Forecasters by the Federal Reserve Bank of Philadelphia has asked forecasters to provide a complete probability distribution for macroeconomic predictions (e.g., changes in gross domestic product, expected future inflation) by indicating probabilities for different bins that constitute the full range of outcomes (Croushore, 1993). Similar methods have been employed in macroeconomic forecasting surveys globally, for example, those conducted by the European Central Bank (Garcia, 2003) and the United Kingdom's National Institute of Economic and Social Research (Wallis, 2004). Beyond professional forecasting, distribution elicitation has been employed in national surveys to gauge consumer expectations regarding a range of economic variables. For example, the Survey of

[1] In this article, we focus on methods of eliciting full belief distributions, but the literature has investigated alternative approaches between eliciting a single data point and eliciting a full distribution. These alternatives include, for instance, eliciting confidence intervals and fractiles. Goldstein and Rothschild (2014) provided a comparison between belief distribution elicitation and other standard formats, including confidence intervals and fractiles, and found that the belief distribution method performs substantially better than other formats at both the individual and aggregate levels.

[2] In baseball, points are called "runs."

Figure 1

Distribution Elicitation Interface Using the Distribution Builder Tool Developed by André (2016; Panel A) and Using Sliders (Panel B)



Note. SPIES= Subjective Probability Interval Estimates. See the online article for the color version of this figure.

Consumer Expectations, launched by the Federal Reserve Bank of New York, has asked respondents to provide entire probability distributions for future inflation rates, home price changes, and earnings changes, which are then used to compute various measures of central tendency and uncertainty (Armantier et al., 2017). Similar approaches have been adopted in other parts of the world, such as in the Survey of Household Income and Wealth carried out by the Bank of Italy.

Given the increasingly widespread use of belief distribution data in both academic, nonacademic, and policymaking contexts, it is important to understand how to elicit that data in a way that best captures participants' true beliefs. But we do not yet know whether different elicitation methods elicit different belief distributions and, if so, whether one method yields greater accuracy.

Not only is the answer to this question practically meaningful, but it is also theoretically relevant. A large literature has explored how logically equivalent elicitation methods can lead respondents to make different choices (Shafir, 1993; Tversky & Kahneman, 1981), provide different point estimates (Kelly & Simmons, 2016; Thomas & Kyung, 2019; Tversky et al., 1988), and generate different interval estimates (Juslin et al., 1999; Klayman et al., 1999; Soll & Klayman, 2004; Teigen & Jørgensen, 2005). Given the many differences between belief distribution elicitation and other question formats, it is worthwhile to investigate whether (and how) logically equivalent elicitation methods may also lead people to generate different distributions of beliefs. To the best of our knowledge, there is currently no research comprehensively investigating how these two methods might induce different response patterns or whether one of them might be superior.

In this article, we report the results of 13 preregistered experiments, 10 reported below and three reported in the Supplemental Materials. In each study, we asked participants to generate belief distributions using either Distribution Builder or Sliders and then assessed the accuracy of these distributions (using benchmarks that we explain below). In some studies, we also manipulated the shape of the correct belief distribution (i.e., right-skewed, symmetric, or left-skewed). We investigated which method yields more accurate belief distributions and why.

Our investigation makes a number of contributions. First, we show that people's belief distributions can be influenced by the method used to elicit them. Second, we show that one popular elicitation method is generally better than another, while also delineating the conditions under which that method is most likely to be superior. Third, by casting light on why one method is superior to another, we gain insight into the factors influencing how belief distributions are constructed. Overall, this work enhances our understanding of the psychology of belief distribution elicitation, enabling us to provide practical guidance about how to better do so going forward.

Research Overview

Do people generate different belief distributions when they use Distribution Builder versus Sliders? And is one method more accurate than the other? We report 10 studies designed to try to answer these questions. Using a variety of stimuli and four different procedures (described in detail below), we found that in most cases, people generate more accurate distributions when they use Distribution Builder than Sliders, with the only (inconsistent) exceptions being when the correct distribution appears right-skewed. Along the way, we shed light on why this happens and report supplemental studies testing the (in)effectiveness of interventions designed to improve the accuracy of Sliders users' distributions.

Before presenting the studies, it is worth describing how we operationalized "accuracy." Indeed, we faced a pretty big and obvious challenge: How do you assess the accuracy of belief distributions when true beliefs are unobservable? Across studies, we used four different accuracy benchmarks, two that were external to participants' beliefs and two that relied on participants' self-reported beliefs. Note that we use the term "accuracy" in a broad way that encompasses the psychometric concepts of both "validity," the extent to which a measure assesses what it is supposed to measure, and "reliability," the extent to which a measure produces consistent results. The first three benchmarks described below assess validity, and the fourth assesses reliability.

In Studies 1–7 and 10, we established external accuracy benchmarks using two procedures. The first procedure resembles the design in Goldstein and Rothschild (2014). In their study, participants first observed, at a fast pace, 100 numbers ranging from 1 to 10. They then indicated the frequency with which each value would occur in another random sample of 100 numbers from the same population. That is, participants saw a distribution of numbers in the first stage and attempted to reproduce that distribution in the second stage. We used a similar "memorize-recall" procedure in Studies 1–6 and 10: Participants were instructed to first observe a sequence of numbers and then to do their best to reproduce the distribution of these numbers. Given that all participants observed the "ground truth" distribution, it is reasonable to infer that belief distributions that are closer to this ground truth are more accurate.[3] We operationalized individual-level accuracy as the absolute deviation of an individual's responses from the true answers (averaged across all bins). Intuitively, this measure captures, on average, how wrong participants' responses are for each bin. So a 2.00 means that, on average, a participant was off by 2.00 for each bin.[4]

We established a different external benchmark in Study 7. Specifically, we moved away from memory tasks and instead assessed the accuracy of participants' subjective probability distributions. The paradigm is similar to one often used in the overconfidence literature (Moore et al., 2015; Soll & Klayman, 2004). Participants were asked four general knowledge questions with straightforward answers (e.g., estimate the age of a person in the photo) and indicated how likely the true answer was to fall into each bin, with the constraint that their probabilities had to sum to 100%. We operationalized accuracy as the probability allocated to the bin containing the correct answer.[5] In general, both external benchmarks assess how well the elicited distributions correspond with empirical facts.

Nevertheless, we recognize that researchers often elicit belief distributions to assess biases or incorrect beliefs. If people's true beliefs are incorrect, then the elicitation method should reflect these inaccuracies. Therefore, in Study 8, we more directly examined how closely the elicited distributions captured participants' true beliefs. In this study, participants made several predictions by reporting their belief distributions. We additionally measured their beliefs in other ways (e.g., asking them to directly state their beliefs) and assessed the degree to which their belief distributions were consistent with these measures. We presume that belief distributions that are more consistent with other, more direct measures of beliefs are more accurate.

Finally, in Study 9, we investigated whether distributions elicited at different time points were consistent with each other. To this end, we used a similar design as Study 8 and asked participants to report two belief distributions for a pair of complementary predictions (i.e., "the percentage of respondents who prefer X over Y" and "the percentage of respondents who prefer Y over X"). We assessed the consistency between these two predictions by measuring the average absolute deviation between the two distributions.

Transparency and Openness

We preregistered every study, and we report all of our measures, manipulations, and exclusions. A detailed breakdown of all exclusions for all studies can be found in Supplemental Material 1. All of our data, materials, and preregistrations are available on ResearchBox at https://researchbox.org/569. This research was approved by the University of Pennsylvania's Institutional Review Board.

[3] Previous research suggests that people encode frequencies with relatively high accuracy (Goldstein & Rothschild, 2014; Hasher & Zacks, 1979, 1984), especially when they experience the stimuli rather than being presented with summary statistics (Camilleri & Newell, 2019; Hogarth & Soyer, 2011).

[4] In exploratory analyses, we also derived Kolmogorov–Smirnov distance as another measure of accuracy. The Kolmogorov–Smirnov distance is defined as the largest absolute difference between the cumulative distribution function of a participant's response and the true distribution function. This measure yields results with the same significance level as our preregistered primary measure. We report these results in Supplemental Material 2 but otherwise do not discuss them further.

[5] In exploratory analyses, we derived another accuracy measure that takes into account the full distribution participants constructed. We multiplied the probability allocated to each bin by the absolute deviation of that bin from the correct bin and summed all these products. This measure captures the aggregate deviation of the elicited distribution from the correct answer. We obtain results with the same significance level as our primary preregistered measure (see Supplemental Material 2 for detailed results).

Studies 1–7

In Studies 1–7, we examined whether Distribution Builder or Sliders elicits more accurate distributions by comparing participants' responses to an external benchmark. Because the seven studies have similar methods and the pattern of results is easier to discern when presented in aggregate, we describe the seven studies all at once.

Method

Participants

We conducted Studies 1–7 online using U.S. participants from Amazon Mechanical Turk (MTurk) and Prolific. We decided in advance to recruit 1,300, 1,000, 1,600, 1,700, 1,700, 2,000, and 1,700 participants, respectively. We preregistered to retain only the first response from Internet Protocol (IP) addresses or worker identifications (IDs) that appeared more than once in our data set. In Studies 1–4 and 6, we preregistered to exclude responses with an average absolute error of zero (because participants with perfect distributions may have been cheating). In Study 5, we preregistered to exclude participants who failed any attention check question and those who provided the same answers to all questions in the first part of the survey. In Studies 1–4, 6, and 7, participants answered one or multiple comprehension check questions about the task instructions prior to being assigned to conditions. Those who failed the comprehension checks twice were not allowed to proceed to the rest of the survey and were thus automatically excluded from our data set. These preregistered exclusions (detailed in Supplemental Material 1) left us with final samples of 1,276, 974, 1,588, 1,573, 1,536, 1,905, and 1,564 participants, respectively. These samples averaged 32–42 years of age; 44%–54% reported their gender as female and 46%–56% as male (see Supplemental Material 1 for a detailed breakdown for each study).

Procedure

We describe the procedures of each of the seven studies below. Studies 1–6 followed a similar "memorize-recall" procedure. We first provide a detailed description of Study 1 and then report the ways in which Studies 2–6 differed from Study 1.

Study 1

Study 1 examined whether Distribution Builder or Sliders prompts users to construct more accurate belief distributions and, because real-world distributions come in different shapes, whether the difference depends on the shape of the true distribution.

We used a procedure similar to Goldstein and Rothschild (2014). Participants learned that they would see a list of 50 randomly selected cars and the miles per gallon (mpg) value for each car. They read that the 50 cars would be presented on the screen five at a time and at a fast pace. Participants were told to focus only on the mpg values and to try to remember as many numbers as possible. Before presenting the stimuli, we asked a one-question comprehension check to verify that they understood the instructions. Participants ($n = 27$) who failed this comprehension check twice were prevented from completing the remainder of the survey.

Figure 2
Example of One Page of Stimuli Presented in Study 1

Dodge Aries Wagon	**26 mpg**
Ford Granada Ghia	**18 mpg**
Chevrolet Impala	**13 mpg**
Honda Civic	**33 mpg**
Datsun B-210	**32 mpg**

Starting on the next page, we presented the brand name and the mpg value for five cars at a time and in a random order, with an exposure duration of 5 s for each page (see Figure 2 for an example).

Participants were randomly assigned to see 50 mpg values from one of three distributions: right-skewed, left-skewed, or symmetric. Panel A of Figure 3 shows the distribution participants saw in these three conditions.

After viewing all 50 numbers, participants were asked to reproduce the distribution of these numbers. Specifically, they were asked to indicate how many cars of each mpg range they saw by allocating 50 numbers into seven labeled bins that covered the entire range: "11–15," "16–20," "21–25," "26–30," "31–35," "36–40," and "41–45." We randomly assigned participants to use either Distribution Builder or Sliders. Those in the Distribution Builder condition saw a graphical interface in which they were asked to place 50 virtual balls into the seven labeled bins by clicking the "+" and "−" buttons below each bin (see Figure 4 Panel A).[6] Those in the Sliders condition saw the same labeled bins with a slider scale next to each one. They were asked to indicate on each slider scale how many cars fell within the mpg range represented by the bin (see Figure 4 Panel B). In both conditions, the interface presented counters displaying how many units the participant had allocated to each bin and the total number of units they had allocated so far. Finally, participants reported their age by entering the number of years and reported their gender by making a choice between "female" and "male."

Studies 2 and 3

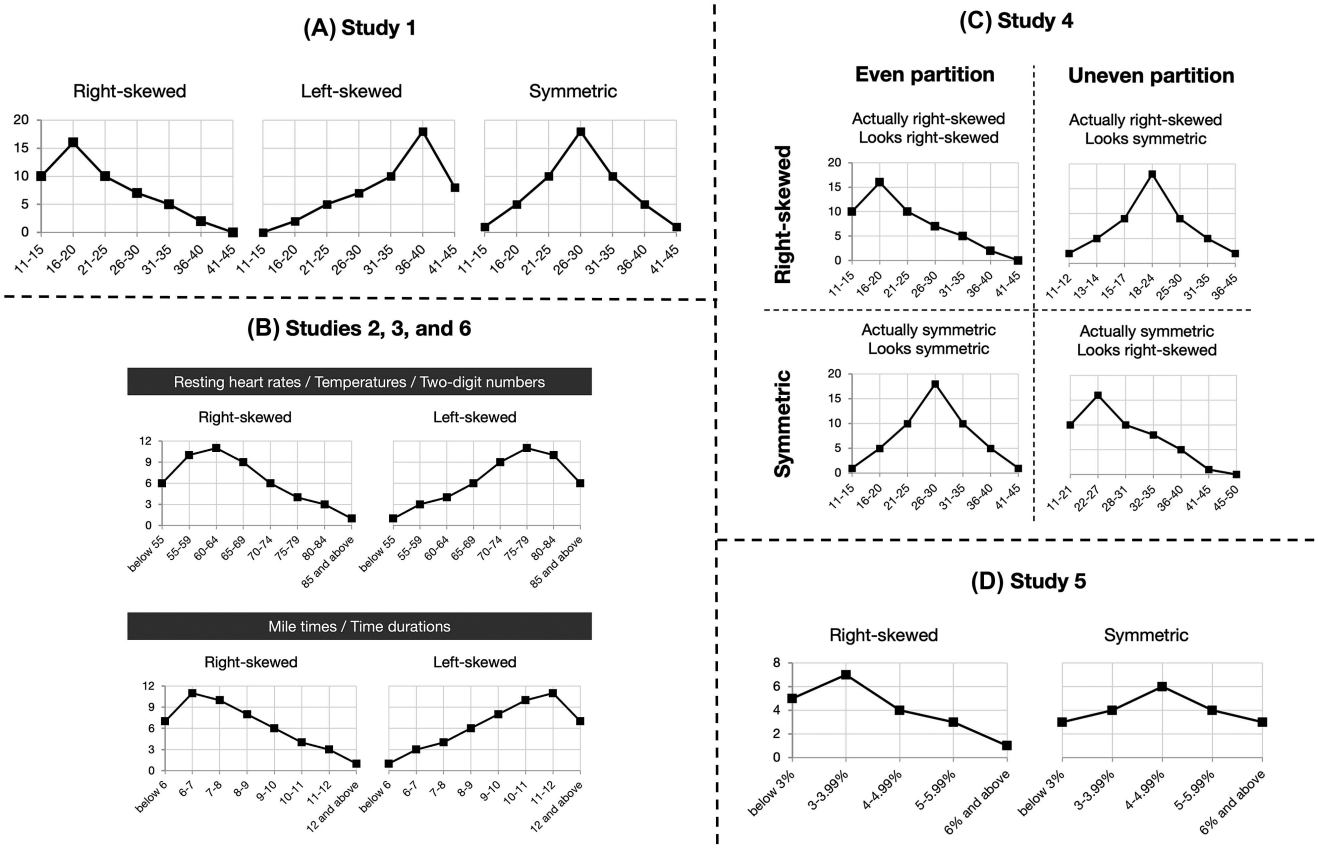
As described below in the Results section, in Study 1, we found that Distribution Builder yielded more accurate distributions than Sliders when the underlying distribution was symmetric or left-skewed but not when the underlying distribution was right-skewed. Studies 2 and 3 aimed to replicate the effect using a within-subject design.

The procedures of Studies 2 and 3 were similar to Study 1, with three notable exceptions. First, all participants completed two rounds of the "memorize-recall" procedure. In Study 2, participants viewed the same set of 50 numbers but were given a different cover story for each round. They read that one set of numbers was the resting heart rate of 50 randomly selected adults and the other set was the air temperature of 50 cities in the world (in Fahrenheit), and we counterbalanced the order in which the two cover stories were presented. The 50 numbers within each round were presented in a randomized order. In Study 3, participants also read these two cover

[6] We used the distBuilder Javascript library developed by André (2016) in all Distribution Builder conditions described in this article.

Figure 3

Distributions of Numbers Presented to Participants (and the Correct Answers) for Each Distribution Condition in Studies 1–6



Note. (A) Distributions presented to participants in Study 1. (B) Distributions presented to participants in Studies 2, 3, and 6. (C) Distributions presented to participants in Study 4. (D) Distributions presented to participants in Study 5. The *x*-axis in each figure represents the bins in the distribution elicitation interface, and the *y*-axis represents the true frequency of numbers in each bin.

stories, but this time, they actually viewed two different sets of 50 numbers. Second, we manipulated elicitation method within-subjects such that participants used Distribution Builder for one round and Sliders for another round, and we counterbalanced the order in which they used the two methods. Third, to simplify the design, we only included right-skewed and left-skewed as the two distribution shape conditions (see Panel B of Figure 3 for details). In Study 2, this manipulation was fully between-subjects, such that participants were assigned to the same distribution shape in both rounds. In Study 3, we manipulated the distribution shape in each round orthogonally, such that each set of numbers was randomly assigned to be from a right-skewed or a left-skewed distribution. This means that across the two rounds, there were in total four distribution shape conditions: right-skewed + right-skewed, right-skewed + left-skewed, left-skewed + right-skewed, and left-skewed + left-skewed.

Study 4

In Study 4, we sought to determine whether the interaction between the elicitation method and the distribution shape was driven by the true shape of the underlying distribution or by how the shape *appeared* in the distribution interface.

Study 4 was identical to Study 1 except for two changes. First, all participants completed two rounds of the "memorize-recall" procedure, where they viewed numbers from a symmetric distribution in one round and numbers from a right-skewed distribution in the other round. The order in which these two rounds were presented was counterbalanced.

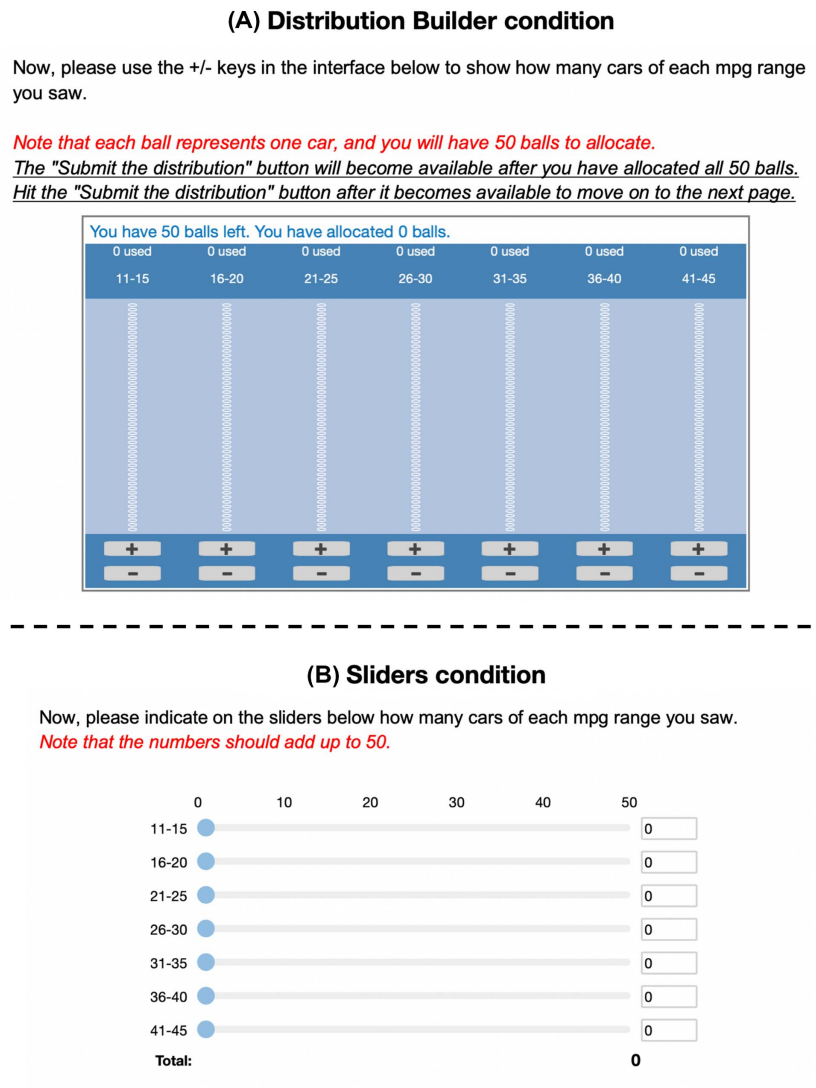
Second, in each round of distribution elicitation, the range in the elicitation interface was randomly assigned to be either evenly or unevenly partitioned. Figure 3 Panel C shows the correct response pattern for each Distribution × Partition condition. Participants in the even partition conditions (see the left column in Figure 3 Panel C) saw the same elicitation interface as in Study 1. In the two uneven partition conditions, we divided the entire range so that the inherently right-skewed distribution had a symmetric-looking answer (top right panel in Figure 3 Panel C) and the inherently symmetric distribution had a right-skewed-looking answer (bottom right panel in Figure 3 Panel C).

Study 5

Study 5 aimed to replicate the results of Studies 1–4 using a different paradigm. Whereas previous studies used an "explicit memorization" procedure, most real-world decision-making contexts

Figure 4

Distribution Elicitation Page Presented in Study 1



Note. (A) Distribution Builder condition. (B) Sliders condition. See the online article for the color version of this figure.

do not require participants to recall information that they explicitly tried to memorize. And so in Study 5, we examined whether the same results would occur when participants were presented with information they were not told to memorize.

Data collection for this study took place on May 26, 2020, a few months into the COVID-19 pandemic. Participants learned that they would answer some questions related to COVID-19 testing rates and reopening policies (as of May 20, 2020) in 20 U.S. states, one state at a time. No information about memorization or recall was mentioned in the instructions, so participants were not expected to try to memorize any of the information.

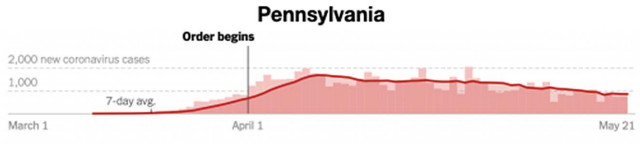
After reading the instructions, participants saw 20 pages, each containing COVID-19 information about one U.S. state. Specifically, we presented a figure showing the trend of confirmed cases, the testing rate, and the reopening policy of this state. On each page, we asked participants two questions regarding their opinion of the testing rate for that particular state, so as to draw their attention to the testing rate, which was the number we later asked them to recall in the subsequent distribution elicitation task. Figure 5 shows an example of one such page presented to participants.

We randomly assigned participants to see 20 states with testing rates from either a right-skewed or a symmetric distribution (see Figure 3 Panel D). We manipulated the shape of the distribution not by changing information of the same states, but by changing the states being presented. This allowed us to present accurate information for each state in the different conditions.

After viewing and rating all 20 states, we asked participants to reproduce the distribution of the testing rates by allocating them into five labeled buckets that covered the entire range: "below 3%," "3%–3.99%," "4%–4.99%," "5%–5.99%," "6% and above."

Figure 5

Example of One Page of Stimuli Presented in Study 5



Testing rate	2.73%
Stay-at-home order	In effect since April 1 and set to expire June 4.
Reopening status	Reopened outdoor and recreation, and retail in regions that meet criteria.

Please indicate to what extent you agree with the following statement.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Pennsylvania has done sufficient testing of COVID-19.	○	○	○	○	○	○	○
Pennsylvania should expand testing of COVID-19 before lifting restrictions on social distancing.	○	○	○	○	○	○	○

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

As in previous studies, participants were randomly assigned to use either Distribution Builder or Sliders.

Study 6

Study 6 examined whether the results of Studies 1–5 are robust to stimuli without meaningful labels. The design was similar to Study 1 except that we also manipulated whether the presented numbers were perceived to be random or meaningful. Participants were randomly assigned to one of eight conditions in a 2 (method: Distribution Builder vs. Sliders) × 2 (distribution shape: right-skewed vs. left-skewed) × 2 (number perception: random vs. meaningful) between-subjects design.

We used the same "explicit memorization" task as in Studies 1–4. Participants learned that they would see two sets of 50 numbers and that they should try to remember as many as they could. In the random number condition, participants read that the two sets of numbers were (a) 50 randomly selected two-digit numbers and (b) 50 randomly selected time durations. In the meaningful number condition, participants learned that the numbers were (a) the resting heart rate of 50 randomly selected adults and (b) the mile time of 50 randomly selected adult male runners.

The next part of the study followed the same procedure as Studies 1–4, in which we first presented 50 numbers and then asked participants to recall the distribution of the numbers they had just seen using either Distribution Builder or Sliders. The same procedure then repeated for the second set of numbers. The order in which the two sets of numbers were presented was counterbalanced. We randomly assigned participants to see the 50 numbers from either a right-skewed distribution or a left-skewed distribution, as shown in Figure 3 Panel B.

Study 7

In the previous studies, we evaluated the performance of the two methods by first showing all participants the same benchmark information and then asking them to reproduce it from memory using either method. Though this procedure permits us to establish a definitive normative benchmark against which individual responses can be objectively evaluated, it does not elicit subjective probability judgments, which is what many researchers try to elicit when using these tasks. We designed Study 7 to address this. Resembling a paradigm commonly used in the overconfidence literature (e.g., Moore et al., 2015; Soll & Klayman, 2004), we asked participants to provide subjective probability distributions for general knowledge questions. Similar to previous studies, we manipulated the elicitation method and the shape of the underlying distribution.

Participants learned that they would be asked about their beliefs about some general knowledge questions. We presented an example question and explained how to interpret responses in the interface participants were assigned to. We asked three comprehension questions to verify that they fully understood these instructions. Participants who failed the comprehension checks twice ($n = 168$) were not allowed to proceed to the rest of the survey.

Upon successful completion of the comprehension checks, participants answered four general knowledge questions (shown in Table 1) by giving their subjective probability distributions with either Distribution Builder or Sliders. The order in which the

Table 1
Questions Used in Study 7

Question domain	Question wording	True answer comes from …	True answer
Weight	Please try to estimate how much the person pictured below weights (in pounds). (A full-length photograph)	The actual weight of the person pictured	165 lb
Age	Please try to estimate how old the person pictured below is. (A half-length portrait)	The actual age of the person pictured.	39 years old
Temperature	Please try to estimate the maximum temperature in New York City on July 6, 2020 (in Fahrenheit).	The average maximum temperature in New York City on July 6 in the past 10 years	89 °F
Walking time	The picture below shows a route from Place A to Place B, with a distance of 1.3 mi. If we randomly select an adult to walk from Place A to Place B at a normal pace without interruption, how many minutes will it take? Please make an estimate. (A Google map screenshot with time estimate and addresses blurred out)	The Google Map time estimate	26 min

questions were presented was counterbalanced. We chose these questions because we thought participants could use common sense to discern the correct range of possible answers and that this range could serve as a reasonable benchmark against which to compare elicited distributions.

Due to the different nature of the external benchmark, our manipulation of the shape of the correct answer was necessarily different. Specifically, we manipulated whether the response pattern would look right-skewed or symmetric by changing whether the correct answer fell at the lower end or in the middle of the range. In the right-skewed condition, the correct answer (as indicated in the last column of Table 1) fell within the second bin (out of seven total bins). In the symmetric condition, the correct answer fell within the fourth bin (out of seven total bins).

Results and Discussion

Analysis Plan

In Studies 1–6, we computed the accuracy of each participant's belief distribution by (a) calculating, for each bin of the distribution, the absolute deviation between the true frequency and the frequency that the participant provided, and then (b) averaging these absolute deviations across all bins. This yielded a measure of the average absolute error for each participant's belief distribution. Intuitively, this measure captures how wrong participants' responses are on average in each bin. Lower numbers represent greater accuracy.

In Study 7, we measured individual accuracy by simply computing the percentage of mass participants allocated to the correct bin (i.e., the second bin in the right-skewed condition or the fourth bin in the symmetric condition). In this case, higher numbers indicate greater accuracy.

In each study, we preregistered to conduct ordinary least squares (OLS) regressions to determine whether Distribution Builder or Sliders elicited more accurate distributions and whether the effect depended on the shape of the distribution and any additional variable of interest (i.e., order of the two methods in Studies 2 and 3, range partition in Study 4, and the meaningfulness of the numbers in Study 6). Specifically, in Studies 1, 5, and 7, we regressed the accuracy measure on (a) the method condition (contrast-coded), (b) the shape condition(s) (contrast-coded), and (c) the interaction(s) between the method condition and the shape condition(s). In Studies 2–4 and 6, we also included in the regression (d) the additional manipulation (contrast-coded; order of the two methods in Studies 2 and 3, range partition in Study 4, and number meaningfulness in Study 6), (e, f) two-way interactions between the additional manipulation and the other two independent variables, and (g) the three-way interaction. In studies in which participants provided more than one belief distribution (Studies 2–4, 6, and 7), the regressions clustered standard errors by participant. In studies with more than one stimulus (Studies 3, 6, and 7), the regressions included fixed effects for stimuli. We discuss any significant interactions below and report the full regression results in Supplemental Material 3.

Accuracy

Table 2 displays the means, and Figure 6 shows the difference between the two method conditions in each study.[7] In Figure 6, a positive and dark bar indicates that Distribution Builder users constructed more accurate distributions, and a negative and white bar indicates that Sliders users constructed more accurate distributions. As shown in Figure 6, Distribution Builder elicited more accurate distributions than Sliders in most cases. Sliders elicited more accurate distributions only in conditions in which the correct answer had a right-skewed shape, but that result was not consistently observed.

The regression results are consistent with this observation. We found a significant interaction between the right-skewed condition and the Sliders condition in Study 1 ($p = .019$), Study 2 ($p < .001$), and Study 4 conditions in which the ranges were evenly partitioned ($p = .025$). This interaction was not significant in Studies 3 and 5–7 ($ps > .204$), where the Distribution Builder was significantly more accurate than Sliders overall, irrespective of the distribution's shape ($p < .001$ in Study 3, $p = .004$ in Study 5, $p < .001$ in Study 6, and $p = .001$ in Study 7).

In addition, in Studies 2 and 3, we found a significant interaction between the elicitation method and the within-subject order ($ps < .001$), such that when Sliders were used in the first round, Slider results were significantly less accurate than Distribution Builder results ($ps < .001$ in Studies 2 and 3), but when Sliders were used in

[7] To facilitate an intuitive understanding of the means presented in Table 2, in Supplemental Table S2.2, we compare them to two benchmarks, the accuracy of completely uniform distributions and the accuracy of completely peaked distributions.

Table 2
Studies 1–7 Results: Raw Means of Accuracy Measure in Distribution Builder Versus Sliders Condition

Study and condition	Distribution Builder		Slider		Paired *t* test
	M	*SD*	*M*	*SD*	
Study 1					
Right-skewed	3.43	1.63	3.16	1.50	$t(418) = 1.76$, $p = .079$
Left-skewed	3.43	1.75	3.66	2.24	$t(428) = -1.19$, $p = .234$
Symmetric	2.85	1.50	3.16	1.93	$t(424) = -1.86$, $p = .064$
Study 2					
Right-skewed	2.51	1.18	2.45	1.22	$t(966) = 0.77$, $p = .443$
Left-skewed	**2.68**	1.40	2.97	1.59	$t(972) = -3.03$, $p = .003$
Study 3					
Right-skewed	**2.42**	1.15	2.59	1.21	$t(1,575) = -2.96$, $p = .003$
Left-skewed	**2.60**	1.37	2.75	1.46	$t(1,585) = -2.02$, $p = .043$
Study 4					
Actually right-skewed/looks right-skewed	3.51	2.07	3.37	1.81	$t(769) = 0.98$, $p = .326$
Actually symmetric/looks symmetric	3.44	1.93	3.65	2.08	$t(769) = -1.43$, $p = .154$
Actually right-skewed/looks symmetric	3.40	1.94	3.51	1.70	$t(792) = -0.88$, $p = .378$
Actually symmetric/looks right-skewed	3.59	1.93	3.53	1.94	$t(785) = 0.50$, $p = .621$
Study 5					
Right-skewed	**1.59**	0.78	1.73	0.95	$t(749) = -2.30$, $p = .022$
Symmetric	1.68	0.85	1.80	1.04	$t(783) = -1.80$, $p = .072$
Study 6					
Right-skewed/meaningful numbers	**2.42**	1.16	2.61	1.27	$t(953) = -2.38$, $p = .017$
Left-skewed/meaningful numbers	**2.55**	1.24	2.86	1.45	$t(954) = -3.57$, $p < .001$
Right-skewed/random numbers	**2.47**	1.28	2.73	1.36	$t(942) = -3.07$, $p = .002$
Left-skewed/random numbers	**2.57**	1.33	2.85	1.57	$t(944) = -2.93$, $p = .004$
Study 7 (all tasks combined)					
Right-skewed	33.05	21.82	31.87	21.56	$t(3,131) = 1.52$, $p = .129$
Symmetric	**29.02**	20.76	26.49	20.89	$t(3,117) = 3.38$, $p < .001$

Note. For Studies 1–6, within each row, boldface indicates that this method generated significantly smaller average absolute error and was significantly more accurate. For Study 7, within each row, boldface indicates that participants using this method allocated significantly more mass to the correct category and was significantly more accurate.

the second round, they were just as accurate or more accurate than Distribution Builder ($p = .020$ in Study 2 and $p = .704$ in Study 3). This is a by-product of the fact that participants seemed to improve over time, and so they were more accurate in the second round than in the first.

The results of Study 4 reveal an additional finding of interest. A significant three-way interaction between elicitation method, distribution shape, and range partition ($p = .009$) indicates that the Method × Distribution Shape interaction was moderated by range partition. When the range of the elicitation interface was unevenly partitioned—such that the inherent shape of the distribution of stimuli did not match the observed shape in the elicitation interface—the pattern of results followed the *observed* shape (i.e., how the distribution appeared in the interface). That is, regardless of the true shape of the underlying distribution, Distribution Builder produced more accurate distributions when the distribution *looked* symmetric, whereas Sliders produced more accurate distributions when the distribution *looked* right-skewed (see Figure 6).

Taken together, we found that when the correct distribution looked left-skewed or symmetric, Distribution Builder always outperformed Sliders, either significantly or directionally. When the correct distribution looked right-skewed, however, Sliders sometimes (but not always) yielded more accurate results. This pattern of results occurred in both explicit (Studies 1–4 and 6) and implicit memorization contexts (Study 5), was robust to whether the numbers were perceived as random or meaningful (Study 6), and generalized to the elicitation of subjective probability distributions (Study 7). We return to why this effect occurs and explore additional measures after further establishing the generalizability of the results in Studies 8 and 9.

Study 8

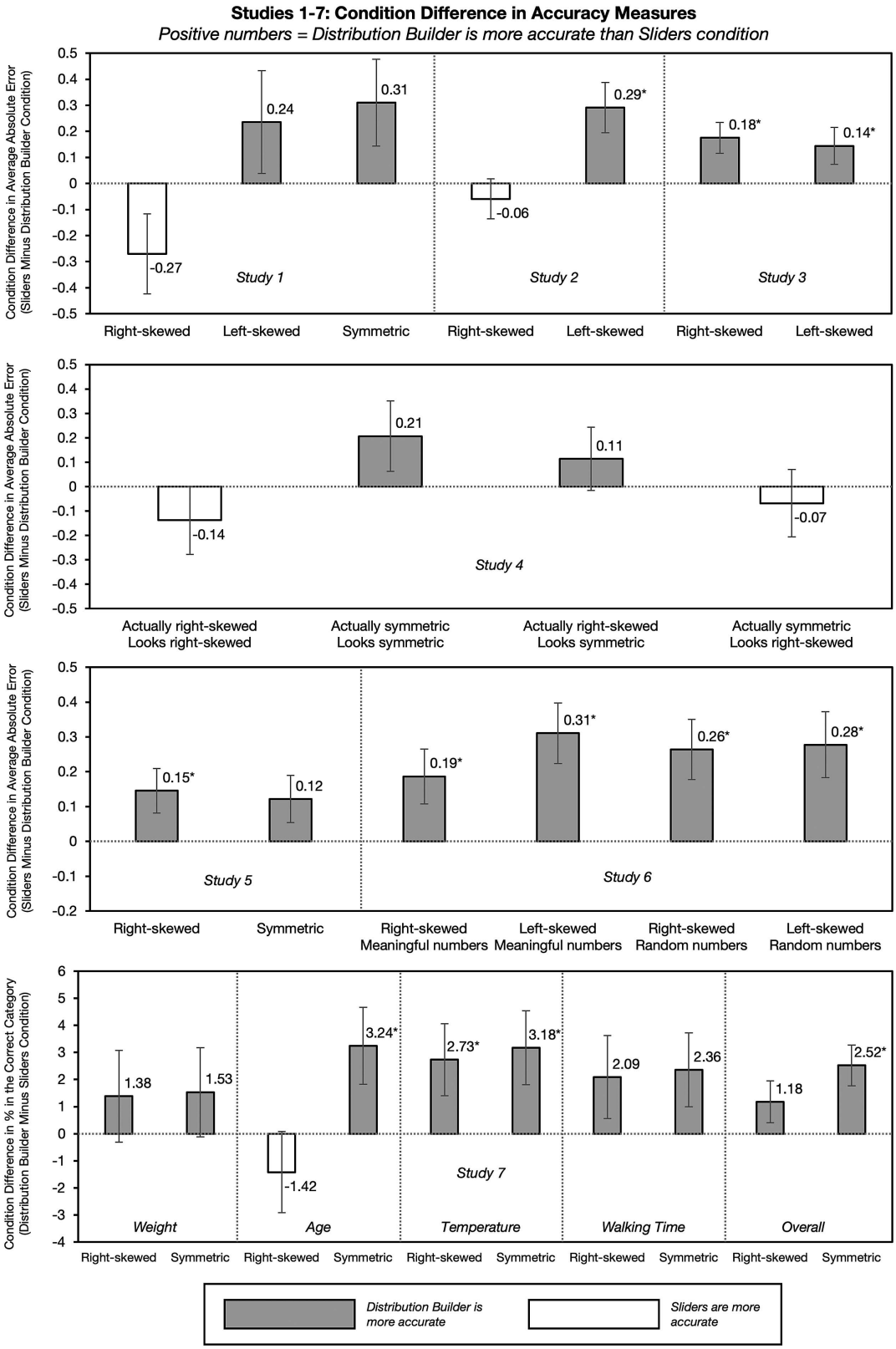
In Studies 1–7, we found that Distribution Builder elicited more accurate distributions than Sliders in most cases, but in these studies, we assessed accuracy by comparing elicited distributions to some ground truth. In Study 8, we explored whether the same findings would emerge when we compared elicited distributions not to empirical facts but to participants' actual beliefs. To accomplish this, in Study 8, we directly measured participants' beliefs and examined which of the two methods elicited distributions that better reflect those beliefs.

Method

Participants

We decided in advance to recruit 1,200 U.S. participants from MTurk. Eight participants who failed an attention check question at the beginning of the survey prior to condition assignment were automatically excluded. We preregistered to retain only the first response from IP addresses or MTurk IDs that appeared more than once in our data set (19 exclusions) and to exclude participants who misreported their MTurk IDs (18 exclusions). This left us with a final sample of 1,162 participants, which averaged

Figure 6
Condition Difference in Accuracy Measures for Studies 1–7



Note. Asterisk indicates a significant difference between Distribution Builder and Sliders ($p < .05$). Error bars represent ±1 *SE*. *SE* = standard error.

39.9 years of age; 555 reported their gender as female (47.8%) and 607 as male (52.2%).

Procedures

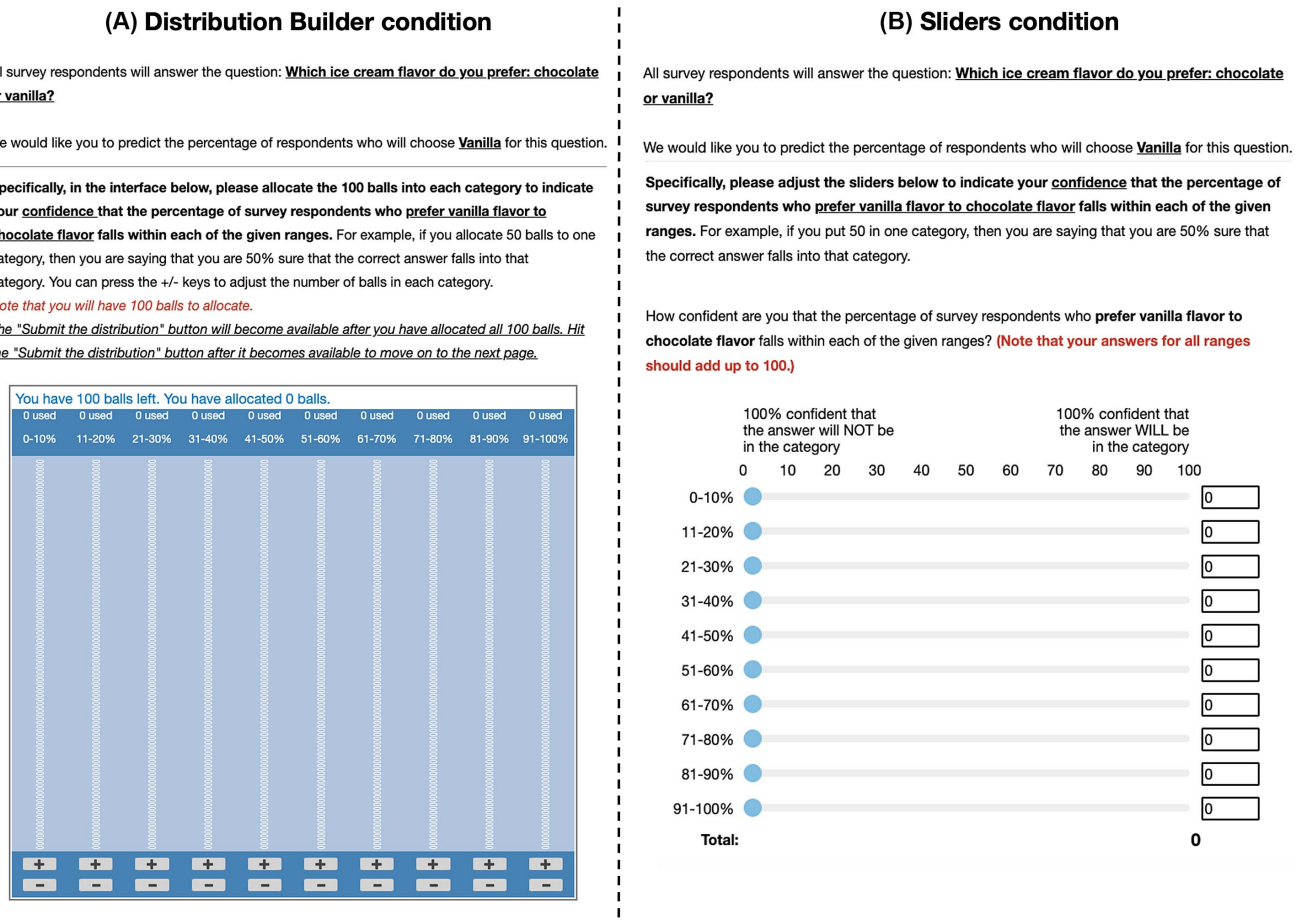
Participants first answered four binary preference/personality questions: (a) "Which ice cream flavor do you prefer: chocolate or vanilla?" (b) "Are you a morning person or a night person?" (c) "Do you prefer the smell of freshly brewed coffee or the smell of freshly baked cookies?" (d) "Would you prefer to be able to see the future or change the past?"

We then randomly selected one of these questions and asked participants to provide a belief distribution that represented their prediction of the percentage of respondents who gave a particular response. Specifically, they were asked to use the Distribution Builder or Sliders interface to indicate how confident they were that the percentage of survey respondents who selected one option over another would fall into each range. So, for example, they indicated how confident they were that the true percentage of people preferring vanilla to chocolate ice cream was within the range of 0%–10%, 11%–20%, 21%–30%, and so on. Within each question, we randomly altered which option participants were predicting. For example, if participants were asked about the ice cream flavor question, they were randomly assigned to predict the percentage of respondents who would choose "chocolate flavor" or "vanilla flavor" for this question. As in previous studies, participants were randomly assigned to provide their belief distributions using either Distribution Builder or Sliders (see Figure 7 for an example).

After providing their belief distributions, all participants indicated which option they thought most participants would choose. For example, for the ice cream flavor prediction, they answered the question, "Do you think more than 50% of survey respondents will choose 'Vanilla' over 'Chocolate'?" (Choice options: "Yes, more survey respondents will choose 'Vanilla.'"; "No, more survey respondents will choose 'Chocolate.'") On the next page, they reported how confident they were in this binary prediction (e.g., "How confident are you that more survey respondents will choose 'Vanilla' over 'Chocolate'?") with possible answers ranging from 50% to 100% in increments of 1%. Finally, they were given an additional $0.40 bonus and were asked to decide how much of this bonus to wager on their binary prediction. At the end of the survey, participants reported their age and gender.

Figure 7

Example of the Distribution Elicitation Page Presented in Study 8



Note. (A) Distribution Builder condition. (B) Sliders condition. See the online article for the color version of this figure.

Results and Discussion

Our goal in this study was to assess how well elicited distributions captured participants' true beliefs. To this end, we first used participants' belief distributions to assess how much confidence they had in their predictions. We did this in a simple way, by computing the percentage of mass they allocated to the half of the range corresponding to their binary prediction. For example, if a participant predicted "chocolate flavor" to be the more popular ice cream flavor, then this measure equaled the amount of mass they allocated to the half of the range corresponding to the chocolate flavor. The more valid participants' belief distributions are, then the more strongly this measure of confidence should correspond with other measures of participants' confidence. We computed two such measures of correspondence.

First, we computed the absolute difference between this belief-distribution measure of confidence and self-reported confidence, namely their response to the question "How confident are you that more survey respondents will [have a certain preference]?" A larger absolute difference indicates that the elicited distribution deviated more from self-reported confidence. An OLS regression[8] revealed that the absolute difference between confidence implied by participants' belief distributions and their self-reported confidence was significantly larger in the Sliders condition ($M = 22.61$, $SD = 18.84$) than in the Distribution Builder condition ($M = 19.08$, $SD = 17.22$), $b = 3.52$, $SE = 1.06$, $p = .001$, indicating that belief distributions in the Sliders condition deviated more from participants' more directly reported beliefs. We replicated this result in Supplemental Study S1.

Second, we examined how consistent elicited distributions were with participants' wager decisions. We preregistered to use OLS to regress the confidence implied by participants' belief distribution on (a) the Slider condition (contrast-coded), (b) the wager amount (mean-centered), and (c) the interaction between (a) and (b), including fixed effects for the prediction question. The interaction term indicates whether the correlation between confidence implied by the belief distribution and the wager amount significantly differed across the two conditions. The interaction term was not significant ($b = -0.13$, $SE = 0.11$, $p = .235$). The results were directionally consistent with those from our previous analysis, as participants' wager amounts significantly predicted the confidence implied by participants' belief distributions in the Distribution Builder condition, $b = 0.23$, $SE = 0.07$, $p = .001$, but not in the Sliders condition, $b = 0.10$, $SE = 0.08$, $p = .213$. We obtained similar results in Supplemental Study S1 (see Supplemental Material 4) and in one unpublished study, but we did not replicate this result in another unpublished study,[9] in which the correlation between wager amounts and the confidence implied by participants' belief distributions was the same ($r = .11$) in both the Distribution Builder and Sliders conditions. We speculate that these weak results may be attributable to the fact that wager decisions do not purely reflect participants' beliefs. For example, they may partially (or largely) reflect participants' risk preferences or their fondness for gambling.

In exploratory analyses, we also derived the proportion of participants who did not allocate more mass to the half of the distribution corresponding to their prediction—in other words, those who produced belief distributions that actually contradicted their binary prediction. This proportion was significantly higher among Sliders users ($M = 28.10\%$, $SD = 0.45$) than Distribution Builder users ($M = 21.99\%$, $SD = 0.41$), $b = 0.06$, $SE = 0.03$, $p = .015$.[10]

These results suggest that the distributions elicited by Distribution Builder more accurately reflected participants' beliefs than did those elicited by Sliders.

Study 9

In Studies 1–8, we found converging evidence that Distribution Builder elicited more accurate distributions than Sliders, both when compared to a ground truth and when compared to people's own beliefs. As an additional test, Study 9 examined whether the two elicitation methods lead to differences in the reliability of elicited belief distributions, that is, the extent to which people provide consistent belief distributions at different time points. We used a paradigm similar to that of Study 8, where participants provided their subjective belief distributions for preference prediction questions. But in this study, participants provided their belief distributions for a pair of complementary predictions—for example, "the percentage of participants who prefer ice cream over frozen yogurt" and "the percentage of participants who prefer frozen yogurt over ice cream"—and we assessed the consistency between these two belief distributions.

Method

Participants

We decided in advance to recruit 1,200 U.S. participants from Connect (CloudResearch). Ten participants who failed an attention check question at the beginning of the survey prior to condition assignment were automatically excluded. We preregistered to retain only the first response from IP addresses or participant IDs that appeared more than once in our data set (12 exclusions) and to exclude participants who misreported their participant IDs (four exclusions). This left us with a final sample of 1,185 participants, which averaged 39.1 years of age; 569 participants reported their gender as female (48.0%) and 616 as male (52.0%).

Procedures

As in Study 8, participants first answered four binary preference questions: (a) "Do you prefer ice cream or frozen yogurt?" (b) "Do you prefer cheeseburger or pizza?" (c) "Do you prefer the feeling of sand between your toes or the feeling of grass under your feet?" (d) "Do you prefer the feeling of a cool breeze on a hot day or the feeling of a warm blanket on a cold day?" We then randomly selected one of these questions and asked participants to provide a belief distribution that represented their prediction of the percentage of respondents who gave a particular response. Participants were

[8] We regressed the absolute difference between these two measures on the Sliders condition (contrast-coded) and included fixed effects for the prediction question.

[9] These unpublished studies were designed to investigate a different, albeit related, hypothesis. Contact the first author for details.

[10] When we restricted our primary analyses to the remaining participants ($n = 871$), Sliders still led to a larger deviation between the confidence implied by participants' belief distributions and their self-reported confidence ($M = 16.58$, $SD = 12.79$) than Distribution Builder ($M = 14.81$, $SD = 12.44$), $b = 1.80$, $SE = 0.86$, $p = .036$.

randomly assigned to use the Distribution Builder or Sliders interface for this prediction.

Different from Study 8, participants completed the belief distribution task twice, once where they predicted the preference of Option 1 over Option 2 (e.g., % of participants who prefer ice cream over frozen yogurt) and once where they predicted the preference of Option 2 over Option 1 (e.g., % of participants who prefer frozen yogurt over ice cream). Participants completed these two belief distributions in a randomized order and completed a filler belief distribution in between, where they predicted another item randomly selected from the remaining three questions.

Results and Discussion

Our measure of interest was the consistency between these two belief distributions. We recoded the second belief distribution such that the bins matched those in the first belief distribution. For example, if a participant indicated that they were 5% confident that 91%–100% of participants prefer frozen yogurt over ice cream, we recoded that as they were 5% confident that 0%–10% of participants prefer ice cream over frozen yogurt. Then, we calculated the average absolute difference between the first distribution and the recoded second distribution the same way as in Studies 1–6. That is, we first calculated, for each bin, the absolute difference in the bin mass between the two distributions and then averaged the absolute differences across all bins. Intuitively, this measure captures how different, on average, the responses in each bin were. Lower numbers represent greater consistency between the two distributions.

We preregistered to use OLS to regress the average absolute difference on the Sliders condition (contrast-coded), including fixed effects for the prediction question. Participants in the Sliders condition produced distributions with greater differences (M = 10.46, SD = 6.51) and thus less consistency than those in the Distribution Builder condition (M = 8.80, SD = 5.95), b = 1.66, SE = .36, $p < .001$. In percentage terms, Sliders elicited distributions that were, on average, 18.9% less consistent than distributions elicited by the Distribution Builder. In other words, the distributions elicited by Distribution Builder were not only more valid but also more reliable.

The Starting Bin in Constructing the Distribution

Thus far, we found that Distribution Builder users constructed more accurate distributions than Sliders users, especially when the underlying distributions were not right-skewed. But why is that? Readers who are looking for an exhaustive answer to that question are, like us, going to be somewhat disappointed. But our analyses of some exploratory measures led us to uncover at least some of the reasons why these two different elicitation methods might produce different responses. Two measures are potentially revealing.

The first is a measure that tracks the percentage of time participants started with the first bin in the interface when providing their belief distributions. In Studies 2–9, we tracked the order in which each bin was first activated on the distribution elicitation page[11] and we consistently found that significantly more participants in the Sliders condition started from the first bin (see the top half of Table 3). And if Sliders users were more likely to start from the first bin, were they more likely to put more mass into that bin? Though the size of the effect varies across studies and conditions, the bottom half of Table 3 does show that that seems to be the case.

Of course, we cannot yet say whether starting with the first bin *causes* participants to put more mass into that bin. But the fact that Sliders users do tend to put more mass in that first bin means that they should often do better when the underlying distribution is right-skewed—since right-skewed distributions have a lot of mass in the first few bins—and worse when the underlying distribution is not right-skewed—since non-right-skewed distributions have little mass in the first few bins. And that is consistent with the findings we have presented thus far.

Study 10

If the tendency to start with the first bin does cause people to put more mass in that bin, then reversing the order of the bins presented to Sliders users—so that the highest bin is presented first and the lowest bin is presented last—should cause them to be more likely to start from the *highest* bin and put more mass in that bin. As a result, these *Reverse* Sliders users should attain lower accuracy when the true distribution is inherently right-skewed (while appearing left-skewed) and higher accuracy when the true distribution is inherently left-skewed (while appearing right-skewed). We tested this in Study 10.

Method

Participants

We decided in advance to recruit 2,000 U.S. participants from Prolific. Participants (n = 248) who failed the comprehension check question twice at the beginning of the survey prior to being assigned to conditions were automatically excluded. We preregistered to retain only the first response from IP addresses or Prolific IDs that appeared more than once in our data set (186 exclusions) and to exclude responses with an average absolute error of zero (three exclusions). This left us with a final sample of 1,787 participants, which averaged 35.8 years of age; 885 reported their gender as female (49.5%) and 902 as male (50.5%).

Procedure

Participants were randomly assigned to one of six conditions in a 3 (method: Distribution Builder vs. Regular Sliders vs. Reverse Sliders) × 2 (distribution shape: right-skewed vs. left-skewed) between-subjects design.

Study 10 used an explicit memorization task that was very similar to the task we used in Studies 1–4 and 6. Participants learned that they would see the danceability value[12] of 50 songs and that they should try to remember as many of these numbers as they could.

[11] We obtained the order data by first tracking the timestamp at which participants first activated each bin on the elicitation interface and then sorting the timestamp values.

[12] All songs in our dataset came from the annual Top Spotify Tracks playlists from 2010 to 2019. Spotify scores every track on the platform on different audio features, including danceability. According to Spotify, "danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity" and ranges from 0 to 100.

Table 3
Exploratory Measures for Studies 1–9

Study and condition	Distribution Builder	Sliders	Condition difference (Sliders − Distribution Builder)
	% of participants who started from the first bin		
Study 2			
Right-skewed	57.1%	**66.7%**	9.6%**
Left-skewed	39.8%	**55.4%**	15.6%***
Study 3			
Right-skewed	58.3%	61.0%	2.8%
Left-skewed	39.5%	**48.8%**	9.3%***
Study 4			
Looks right-skewed	53.8%	**69.8%**	16.0%***
Looks symmetric	49.0%	**64.1%**	15.1%***
Study 5[a]			
Right-skewed	49.2%	**58.8%**	9.6%**
Symmetric	50.0%	**63.1%**	13.1%***
Study 6[a]			
Right-skewed	45.8%	**55.1%**	9.3%***
Left-skewed	39.2%	**52.2%**	13.0%***
Study 7[a]			
Right-skewed	38.5%	**58.2%**	19.7%***
Symmetric	12.0%	**31.6%**	19.6%***
Study 8			
All data	16.8%	**51.7%**	34.9%***
Study 9			
All data	11.5%	**38.0%**	26.5%***
	% of mass allocated to the first bin		
Study 1			
Right-skewed	11.5%	**14.2%**	2.7%**
Left-skewed	2.2%	3.0%	0.8%
Symmetric	3.4%	4.6%	1.2%
Study 2			
Right-skewed	7.0%	7.5%	0.5%
Left-skewed	3.7%	4.2%	0.6%
Study 3			
Right-skewed	7.3%	7.0%	−0.3%
Left-skewed	3.3%	3.1%	−0.2%
Study 4			
Looks right-skewed	11.2%	**12.7%**	1.5%**
Looks symmetric	7.8%	**8.9%**	1.1%*
Study 5			
Right-skewed	25.2%	**28.6%**	3.4%**
Symmetric	15.3%	**17.6%**	2.3%**
Study 6			
Right-skewed	8.3%	8.1%	−0.2%
Left-skewed	4.0%	4.1%	0.1%
Study 7			
Right-skewed	20.3%	**24.4%**	4.0%***
Symmetric	3.9%	**5.9%**	1.9%***
Study 8			
All data	3.0%	**6.4%**	3.3%***
Study 9			
All data	3.2%	**7.0%**	3.7%***

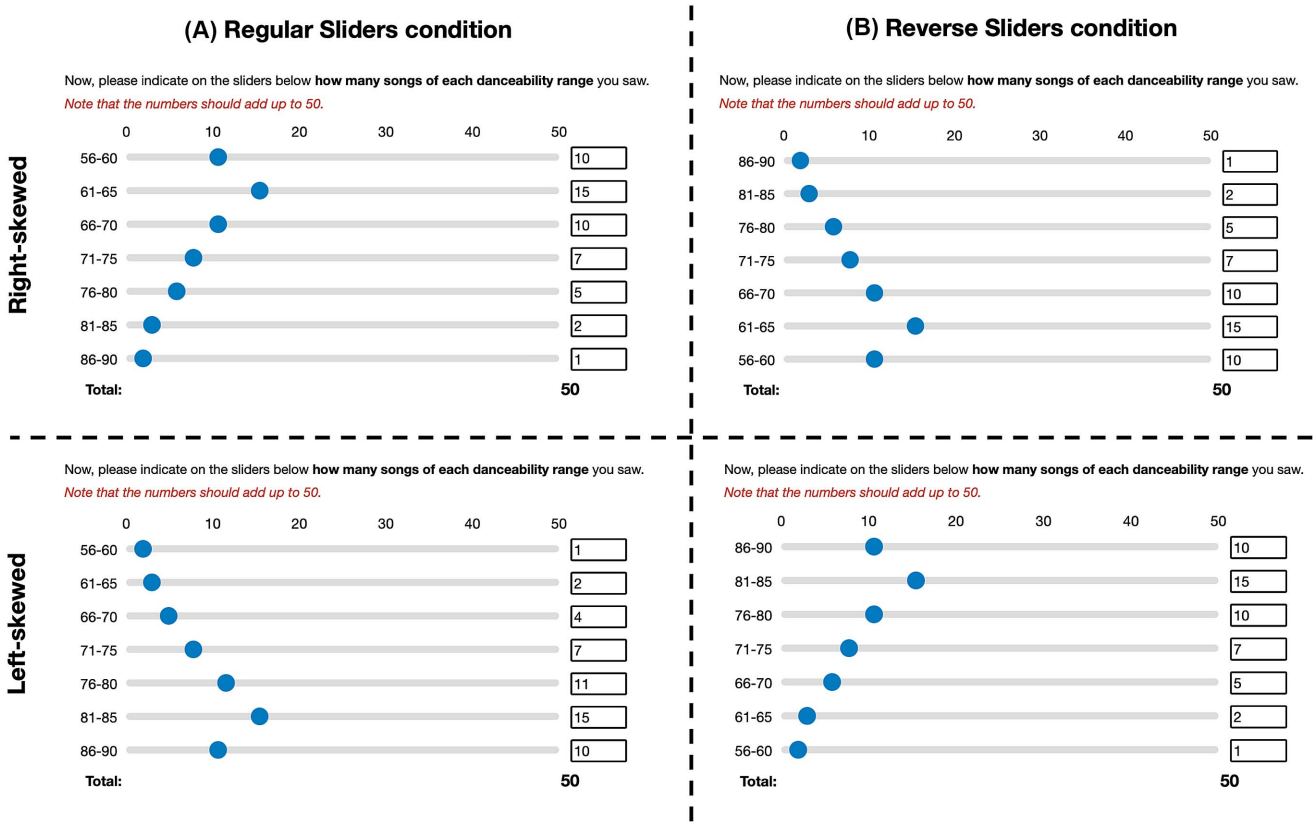
Note. Within each row, boldface indicates that this percentage is significantly higher than the adjacent percentage.
[a] This measure was preregistered.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Upon passing a comprehension check, participants observed 50 numbers, with 5 numbers presented on each screen for 5s.

Similar to previous studies, we randomly assigned participants to see the 50 numbers from either a right-skewed distribution or a left-skewed distribution. We also randomly assigned participants to try to reproduce the distribution they saw using Distribution Builder, Regular Sliders, or Reverse Sliders. The instructions and interfaces for the first two method conditions were the same as in previous studies. In the Reverse Sliders condition, the sliders were reversed so that the highest bin was at the top and the lowest bin was at the bottom (see Figure 8 Panel B). Therefore, correct distributions in the Reverse Sliders condition would have been left-skewed-looking if

Figure 8

The Belief Distribution Elicitation Page (and the Correct Answer in Each Distribution Condition) Presented in Study 10



Note. (A) Regular sliders condition. (B) Reverse sliders condition. See the online article for the color version of this figure.

participants observed a right-skewed distribution and right-skewed-looking if participants observed a left-skewed distribution. Finally, participants reported their age and gender.

Results and Discussion

As in our previous memorization-task studies, we used average absolute error as the measure of accuracy. Figure 9 displays the results. First, comparing Regular Sliders to Distribution Builder, we replicated our previous results: Distribution Builder users generated significantly more accurate distributions when the underlying distribution was left-skewed ($p = .002$) but not when the underlying distribution was right-skewed ($p = .762$; interaction: $b = -0.43$, $SE = 0.21$, $p = .035$). Importantly, this effect reversed when we compared Reverse Sliders to Distribution Builder: Distribution Builder users provided significantly more accurate distributions when the underlying distribution was right-skewed (but left-skewed-looking for Reverse Sliders; $p = .016$) but not when the underlying distribution was left-skewed (but right-skewed-looking for Reverse Sliders; $p = .869$; interaction: $b = 0.40$, $SE = 0.21$, $p = .052$).

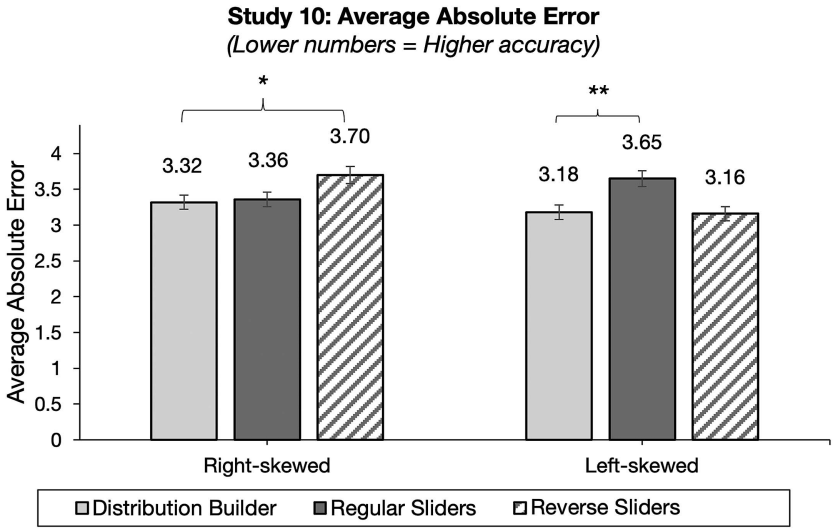
These results suggest that changing the order in which the bins are presented in the Sliders interface can change which bin Sliders users start with, and consequently the mass assigned to that bin. To see this more directly, Table 4 presents the results of exploratory analyses suggesting that reversing the order of the bins causes Sliders users to (a) start with the highest bin more often and assign more mass to that bin and (b) start with the lowest bin less often and assign (slightly) less mass to that bin.

All told, these results suggest that Sliders users are more likely than Distribution Builder users to start from the first bin in the interface and to therefore assign more mass to that bin. When the true distribution does not contain a lot of mass in the first bin—as is the case when the true distribution is symmetric around a middle bin, or when the true distribution is left-skewed—then this results in more erroneous distributions.

General Discussion

Belief elicitation is fundamental to many fields of social science research. In recent years, researchers in psychology and adjacent fields have increasingly adopted the approach of eliciting people's entire belief distributions in an attempt to get a more complete picture of people's beliefs. The current research shows that people generate different belief distributions depending on which (popular) method they are using: Distribution Builder or Sliders. Across 10 studies, we found that Distribution Builder outperformed Sliders in most cases. This advantage held across different domains and different types of tasks: explicit memorization tasks

Figure 9
Study 10 Accuracy Results



Note. Error bars represent ±1 *SE*. *SE* = standard error.
* $p < .05$. ** $p < .01$.

(Studies 1–4, 6, and 10), implicit memorization tasks (Study 5), and subjective probability judgments (Studies 7–9).[13]

Why do Sliders elicit less accurate distributions? Our studies provide converging evidence for one explanation: Sliders users are more likely to start with the first bin and assign more mass to that bin, resulting in overly right-skewed distributions (Table 3). Because of this, Sliders users (sometimes) produce more accurate distributions when true distributions are right-skewed, and they otherwise produce less accurate distributions. The most direct evidence for this process comes from Study 10, in which we found that reversing the order of the slider scales reversed our typical effect. This time, Sliders users were more accurate for *left-skewed* (but right-skewed-looking) distributions than for right-skewed (but left-skewed-looking) distributions.

Figure 10 presents the effect sizes of the difference between the two methods. On average, the effect sizes were not particularly large, though they are of a similar magnitude to the documented effects of many behavioral interventions (Mertens et al., 2022; Szaszi et al., 2022). Nevertheless, it is worth noting that the difference between the two methods was larger when we used participants' own beliefs as the benchmark and provided no ground truth prior to the elicitation ($d = 0.20$ in Study 8 and $d = 0.27$ in Study 9), a design that is closer to how researchers and practitioners typically elicit beliefs. Though modest in absolute terms, these differences may have a substantial influence when elicited distributions are used as inputs for economic and policy decisions with far-reaching implications.

A large literature suggests that logically equivalent response formats can generate different responses to simple questions (e.g., Juslin et al., 1999; Kelly & Simmons, 2016; Klayman et al., 1999; Moon & Nelson, 2020; Shafir, 1993; Soll & Klayman, 2004; Teigen & Jørgensen, 2005; Thomas & Kyung, 2019; Tversky & Kahneman, 1981; Tversky et al., 1988), and researchers have identified many different reasons why. For example, response formats may change people's reference points (e.g., Tversky & Kahneman, 1981) or alter the salience of different information, options, or stimuli (e.g., Kelly & Simmons, 2016). Different response formats may also invite respondents to incorporate different kinds of information into their responses, as when respondents feel more inclined to incorporate market prices into willingness-to-pay measures than into other measures of preference (e.g., Evangelidis et al., 2022; Moon & Nelson, 2020). And different response formats may generate different responses for more mechanical reasons, such as scale compatibility, the idea that people give more weight to stimulus attributes that are compatible with the response mode (Tversky et al., 1988).

Adding to this work, our research shows that two logically identical elicitation methods can lead people to generate different belief distributions. This suggests that, just as with other stated beliefs and preferences, people may not have perfectly stable belief distributions and that they at least partially construct them when elicited. This is important because constructing a belief distribution is inherently different from responding to a single question. It is more dynamic, involves multiple steps, and potentially leaves room for more opportunities to be influenced by contextual cues.

An observation of the two methods reveals several differences that one might expect to influence the process of belief distribution construction. Perhaps most notably, Distribution Builder prompts users to create a vertical-looking histogram, while the Sliders prompt users to create a horizontal-looking histogram. Another obvious difference is that Distribution Builder requires more

[13] The current work examines the absolute accuracy of elicited distributions and speaks to the psychometric properties of validity and reliability. We do not have evidence to speak to the sensitivity of the two methods, that is, whether the two methods are sensitive to manipulations that should change the elicited distributions.

Table 4
Exploratory Measures for Study 10

Measure and condition	Distribution Builder	Regular Sliders	Reverse Sliders	Effect of the Sliders condition (vs. Distribution Builder condition)	Effect of the Reverse Sliders condition (vs. Distribution Builder condition)
% of participants who started from the lowest bin (56–60)					
Right-skewed	53.66%	58.64%	15.13%	$b = 0.05$, $SE = 0.04$, $p = .226$	$b = -0.39$, $SE = 0.04$, $p < .001$
Left-skewed	52.13%	57.98%	17.65%	$b = 0.06$, $SE = 0.04$, $p = .146$	$b = -0.34$, $SE = 0.04$, $p < .001$
% of participants who started from the highest bin (86–90)					
Right-skewed	6.62%	9.15%	49.67%	$b = 0.03$, $SE = 0.02$, $p = .258$	$b = 0.43$, $SE = 0.03$, $p < .001$
Left-skewed	13.77%	13.03%	64.36%	$b = -0.01$, $SE = 0.03$, $p = .788$	$b = 0.51$, $SE = 0.03$, $p < .001$
% of mass allocated to the lowest bin (56–60)					
Right-skewed	14.20%	15.51%	13.24%	$b = 0.01$, $SE = 0.01$, $p = .095$	$b = -0.01$, $SE = 0.01$, $p = .168$
Left-skewed	3.65%	5.79%	3.63%	$b = 0.02$, $SE = 0.01$, $p < .001$	$b < 0.001$, $SE = 0.004$, $p = .944$
% of mass allocated to the highest bin (86–90)					
Right-skewed	3.09%	3.21%	5.06%	$b = 0.001$, $SE = 0.01$, $p = .815$	$b = 0.02$, $SE = 0.01$, $p = .001$
Left-skewed	15.69%	14.73%	18.93%	$b = -0.01$, $SE = 0.01$, $p = .264$	$b = 0.03$, $SE = 0.01$, $p < .001$

Note. SE = standard error.

granular and precise inputs than Sliders, as Distribution Builder users need to click the button one at a time in order to allocate their responses, whereas Sliders users can directly drag the sliding bar to their desired value. This means that Distribution Builder requires more effort and longer completion time than Sliders, which may prompt more deliberation in the meantime. We examined these two possibilities in subsequent studies.

First, people may be more likely to start from the lower bins when the belief distribution interface elicits a horizontal-looking histogram (as is the case with Sliders) than a vertical-looking histogram (as is the case with Distribution Builder). To test this, we conducted Supplemental Study S1 (described in Supplemental Material 4), in which we manipulated whether participants were randomly assigned to use Distribution Builder, Regular (Horizontal) Sliders, or Vertical Sliders (i.e., a vertical version of the sliders interface).[14] We used the same study procedure as Study 8, in which participants constructed belief distributions for certain prediction questions and then indicated their confidence using two other, more direct measures. Replicating Study 8, we found a bigger gap between the confidence implied by participants' belief distributions and their self-reported confidence among participants who used Sliders than among those who used Distribution Builder. More importantly, this difference did not diminish—and in fact directionally increased—among those who used Vertical Sliders.[15] This study suggests that the difference in how the Sliders versus Distribution Builder interface is oriented is not responsible for the fact that Distribution Builder tends to elicit more accurate distributions.

Another possible explanation arises from the fact that completing the Distribution Builder task involves clicking the +/− button for each input, which presumably requires more time and effort than simply moving the slider scales. This additional effort requirement might cause Distribution Builder users to be more careful or deliberate. To investigate whether differential deliberation may be responsible for the accuracy gap between Sliders users and Distribution Builder users, we conducted Supplemental Study S2 (described in Supplemental Material 5), in which we asked participants to try to reproduce a distribution of numbers that they were asked to remember. Participants were randomly assigned to use Sliders or Distribution Builder, and they did this at two points in time, using the same interface. At Time 1, participants provided their distributions as usual. Then, at Time 2, participants learned that they would have the opportunity to revise their previous answers and could earn a bonus payment of up to $1, depending on the accuracy of their final submission. On the next page, participants saw their previous responses prepopulated in the elicitation interface and were allowed to make any changes before submitting their second response.

If Distribution Builder produces more accurate distributions because it produces more careful responses, then incentivizing participants to be more careful should increase the accuracy of Sliders users' distributions more than it increases the accuracy of Distribution Builder users' distributions, thereby decreasing the accuracy gap between them. But that is not what we found. Instead, we found that incentivizing participants for accuracy increased the accuracy of users' distributions, and that accuracy gain was the same for those using Sliders and those using Distribution Builder. It makes sense that encouraging participants to think more carefully would increase the accuracy of their distributions. Nevertheless, we do not have evidence to preclude or support the possibility that deliberation (or the lack thereof) led participants to construct distributions differently in the first place. More research is needed.

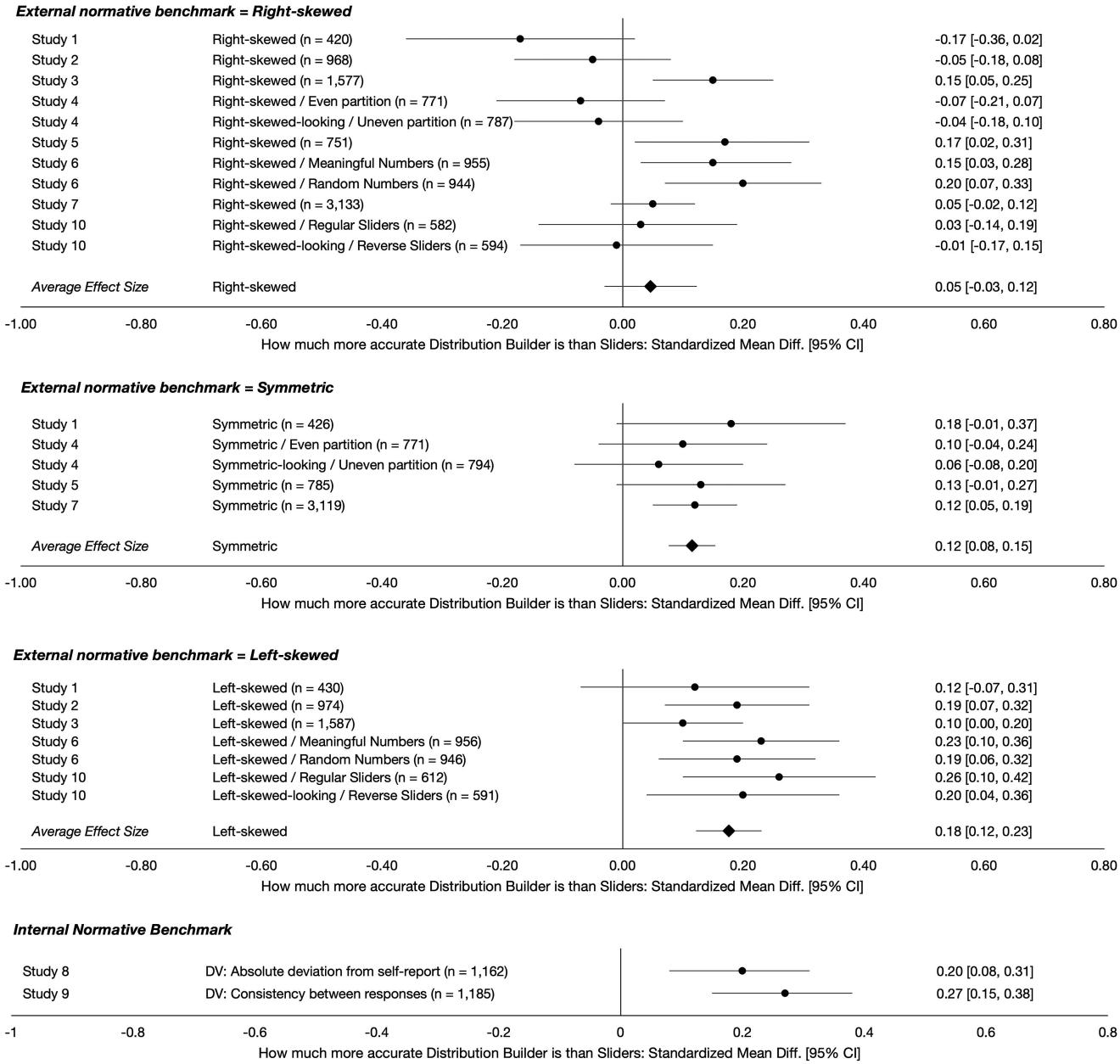
More research is also needed on how to make Sliders users' distributions as accurate as those elicited by Distributed Builder. Because Sliders users tend to start with the first few bins in the interface, and because that seems to lead them to put too much mass in those bins, we thought that instructing them to start with

[14] There are two ways one could imagine testing this question: rotating the orientation of the Sliders interface or rotating the orientation of the Distribution Builder interface. We chose to go with the former approach because we were primarily interested in a manipulation that may improve Sliders (rather than decrease the performance of Distribution Builder).

[15] We did not have data on whether participants start with the first bin in the Vertical Sliders condition, and thus cannot speak to whether the difference (or lack thereof) in the starting bin might explain these results. Nevertheless, participants using the Vertical Sliders also tend to allocate more mass into the first bin compared to the Distribution Builder condition (see Supplemental Material 4 for details), which is consistent with the results presented thus far and suggests that the orientation on its own cannot explain the difference between Distribution Builder and Sliders.

Figure 10

Forest Plot of the Effect Sizes Between the Distribution Builder Condition and the Sliders Condition (Cohen's d, Standardized Mean Difference)



Note. A positive sign means that Distribution Builder elicited more accurate distributions than Sliders; a negative sign means that Sliders elicited more accurate distributions than Distribution Builder. CI = confidence interval.

a different bin might result in more accurate distributions. We tested this possibility in Supplemental Study S3 (described in Supplemental Material 6).

In this study, participants read that they would observe the resting heart rate of 50 randomly selected adults, presented in rapid succession, and from a left-skewed distribution. We manipulated whether participants used Sliders versus Distribution Builder to construct their belief distributions, and whether participants were instructed to start from the "modal" bin. In the Control (i.e., no order instructions) condition, the procedure was exactly the same as in other studies: Participants proceeded to the elicitation interface after observing the 50 numbers. In the Order instructions condition, participants were first shown a page with all possible bins and chose the one they thought contained the most numbers they saw. Then, on a separate page, they saw a screenshot of the Distribution Builder or Sliders interface with the chosen bin circled and were asked to start

with that bin when constructing the distribution on the following page.

Although these order instructions were successful at getting participants to start with the "modal" bin, they actually exerted a backfire effect, whereby participants receiving these instructions provided distributions that were much less accurate, regardless of which interface they used. It seems that instructing participants to start with the "modal" bin may have led them to allocate too much mass to that bin, which in turn decreased their accuracy.

We are still on the hunt for an intervention that can both (a) increase the accuracy of distributions elicited by Sliders and (b) make Sliders users' distributions as reliably accurate as those elicited by Distribution Builder. In the meantime, our advice is to use Distribution Builder.

Constraints on Generality

Our experiments recruited U.S. online participants from the Amazon Mechanical Turk, Prolific, and CloudResearch platforms. While many researchers who elicit belief distributions use a similar participant pool, it is still important to understand whether our findings would generalize to expert samples (who might have more experience with constructing belief distributions using one of the methods), to non-U.S. samples, and to the (large) part of the U.S. population who do not take these online surveys.

References

André, Q. (2016). *distBuilder*. https://quentinandre.github.io/DistributionBuilder/

André, Q., Reinholtz, N., & De Langhe, B. (2022). Can consumers learn price dispersion? Evidence for dispersion spillover across categories. *Journal of Consumer Research*, *48*(5), 756–774. https://doi.org/10.1093/jcr/ucab030

Armantier, O., Topa, G., Van der Klaauw, W., & Zafar, B. (2017). An overview of the survey of consumer expectations. *Economic Policy Review*, *23-2*, 51–72. https://www.newyorkfed.org/medialibrary/media/research/epr/2017/epr_2017_survey-consumer-expectations_armantier.pdf

Bechler, C. J., & Levav, J. (2022). Compatibility effects in the perception of dispersion. *Cognition*, *225*, Article 105166. https://doi.org/10.1016/j.cognition.2022.105166

Bokern, P., Linde, J., Riedl, A., Schmeets, H., & Werner, P. (2021). A survey of risk preference measures and their relation to field behavior. *Netspar Survey Paper*, *58*. https://www.netspar.nl//assets/uploads/P20210720_Netspar-Survey-Paper-176_WEB.pdf

Camilleri, A. R., Cam, M. A., & Hoffmann, R. (2019). Nudges and signposts: The effect of smart defaults and pictographic risk information on retirement saving investment choices. *Journal of Behavioral Decision Making*, *32*(4), 431–449. https://doi.org/10.1002/bdm.2122

Camilleri, A. R., & Newell, B. R. (2019). Better calibration when predicting from experience (rather than description). *Organizational Behavior and Human Decision Processes*, *150*, 62–82. https://doi.org/10.1016/j.obhdp.2018.10.006

Croushore, D. (1993). *Introducing: The survey of professional forecasters*. Business review, Federal Reserve Bank of Philadelphia 3, November/December, 3–15.

Delavande, A., & Rohwedder, S. (2008). Eliciting subjective probabilities in Internet surveys. *Public Opinion Quarterly*, *72*(5), 866–891. https://doi.org/10.1093/poq/nfn062

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Dimant, E. (2023). Beyond average: A method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*, *233*, Article 111417. https://doi.org/10.1016/j.econlet.2023.111417

Dimant, E., Gelfand, M. J., Hochleitner, A., & Sonderegger, S. (2024). Strategic behavior with tight, loose, and polarized norms. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2023.01022

Donkers, B., Lourenço, C., Goldstein, D., & Dellaert, B. (2013). *Building a distribution builder: Design considerations for financial investment and pension decisions* [Netspar design paper 20]. Netspar.

Evangelidis, I., Jung, M., & Moon, A. (2022). *When willingness-to-pay seems irrational: The role of perceived market price*. https://doi.org/10.2139/ssrn.4294247

Garcia, J. A. (2003). *An introduction to the ECB's Survey of Professional Forecasters* [Occasional paper series, No. 8]. European Central Bank.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701. https://doi.org/10.1198/016214505000000105

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, *35*(3), 440–456. https://doi.org/10.1086/589562

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1–14. https://doi.org/10.1017/S1930297500004940

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467–476. https://doi.org/10.1017/S1930297500001637

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*(3), 356–388. https://doi.org/10.1037/0096-3445.108.3.356

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*(12), 1372–1388. https://doi.org/10.1037/0003-066X.39.12.1372

Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020, April). *How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results* [Conference session]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–12).

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General*, *140*(3), 434–463. https://doi.org/10.1037/a0023265

Hu, B., & Simmons, J. P. (2023). Does constructing a belief distribution truly reduce overconfidence? *Journal of Experimental Psychology: General*, *152*(2), 571–589. https://doi.org/10.1037/xge0001291

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1038–1052. https://doi.org/10.1037/0278-7393.25.4.1038

Kelly, T. F., & Simmons, J. P. (2016). When does making detailed predictions make predictions worse? *Journal of Experimental Psychology: General*, *145*(10), 1298–1311. https://doi.org/10.1037/xge0000204

Klayman, J., Soll, J. B., González-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

Leemann, L., Stoetzer, L. F., & Traunmüller, R. (2021). Eliciting beliefs as distributions in online surveys. *Political Analysis*, *29*(4), 541–553. https://doi.org/10.1017/pan.2020.42

Lim, S., Donkers, B., van Dijl, P., & Dellaert, B. G. (2021). Digital customization of consumer investments in multiple funds: Virtual integration

improves risk–return decisions. *Journal of the Academy of Marketing Science*, *49*(4), 723–742. https://doi.org/10.1007/s11747-020-00740-4

Long, A. R., Fernbach, P. M., & De Langhe, B. (2018). Circle of incompetence: Sense of understanding as an improper guide to investment risk. *Journal of Marketing Research*, *55*(4), 474–488. https://doi.org/10.1509/jmr.16.0429

Mannes, A. E., & Moore, D. A. (2013). A behavioral demonstration of overconfidence in judgment. *Psychological Science*, *24*(7), 1190–1197. https://doi.org/10.1177/0956797612470700

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376. https://doi.org/10.1111/j.1468-0262.2004.00537.x

Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1), Article e2107346118. https://doi.org/10.1073/pnas.2107346118

Moon, A., & Nelson, L. D. (2020). The uncertain value of uncertainty: When consumers are unwilling to pay for what they like. *Management Science*, *66*(10), 4686–4702. https://doi.org/10.1287/mnsc.2019.3426

Moore, D. A., Carter, A. B., & Yang, H. H. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes*, *131*, 110–120. https://doi.org/10.1016/j.obhdp.2015.09.003

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Muthukrishna, M., Henrich, J., Toyokawa, W., Hamamura, T., Kameda, T., & Heine, S. J. (2018). Overconfidence is universal? Elicitation of Genuine Overconfidence (EGO) procedure reveals systematic differences across domain, task knowledge, and incentives in four populations. *PLOS ONE*, *13*(8), Article e0202288. https://doi.org/10.1371/journal.pone.0202288

Page, L., & Goldstein, D. G. (2016). Subjective beliefs about the income distribution and preferences for redistribution. *Social Choice and Welfare*, *47*(1), 25–61. https://doi.org/10.1007/s00355-015-0945-9

Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, *12*(1), 29–41. https://doi.org/10.1017/S1930297500005222

Reinholtz, N., Fernbach, P. M., & De Langhe, B. (2021). Do people understand the benefit of diversification? *Management Science*, *67*(12), 7322–7343. https://doi.org/10.1287/mnsc.2020.3893

Ren, Y., & Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management Science*, *59*(11), 2502–2517. https://doi.org/10.1287/mnsc.2013.1715

Robbins, T., & Hemmer, P. (2018). *Lay understanding of illness probability distributions* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society, Madison, WI, United States.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556. https://doi.org/10.3758/BF03197186

Sharpe, W. F., Goldstein, D. G., & Blythe, P. W. (2000). *The distribution builder: A tool for inferring investor preferences*. https://web.stanford.edu/~wfsharpe/art/qpaper/qpaper.html

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

Soll, J. B., Palley, A. B., Klayman, J., & Moore, D. A. (2023). Overconfidence in probability distributions: People know they don't know, but they don't know what to do about it. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2019.00660

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(31), Article e2200732119. https://doi.org/10.1073/pnas.2200732119

Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, *19*(4), 455–475. https://doi.org/10.1002/acp.1085

Thomas, M., & Kyung, E. J. (2019). Slider scale or text box: How response format shapes responses. *The Journal of Consumer Research*, *45*(6), 1274–1293. https://doi.org/10.1093/jcr/ucy057

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*(3), 371–384. https://doi.org/10.1037/0033-295X.95.3.371

Wallis, K. F. (2004). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, *189*, 64–71. https://doi.org/10.1177/002795010418900107

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33). https://doi.org/10.1073/pnas.2302491120

Received March 20, 2023
Revision received July 3, 2024
Accepted July 11, 2024 ■