Overconfidence in Probability Distributions:

People know they don't know but they don't know what to do about it

Forthcoming at *Management Science*

Jack B. Soll

Fuqua School of Business, Duke University, Durham, NC 27708. jsoll@duke.edu

Asa B. Palley

Kelley School of Business, Indiana University, Bloomington, IN 47405. apalley@indiana.edu

Joshua Klayman

The University of Chicago Booth School of Business, Chicago, IL 60637. joshk@uchicago.edu

Don A. Moore

Haas School of Business, University of California, Berkeley, CA 94720. dmoore@haas.berkeley.edu

Author note: Preregistrations, materials, and data are available at https://osf.io/dt7cq/.

Abstract: Overconfidence is pervasive in subjective probability distributions (SPDs). We develop new methods to analyze judgments that entail both a distribution of possible outcomes in a population (aleatory uncertainty) and imperfect knowledge about that distribution (epistemic uncertainty). In four experiments we examine the extent to which subjective probability mass is concentrated in a small portion of the distribution versus spread across all possible outcomes. We find that although SPDs roughly match the concentration of the empirical, aleatory distributions, people's judgments are consistently overconfident because they fail to spread out probability mass to account for their own epistemic uncertainty about the location and shape of the distribution. Although people are aware of this lack of knowledge, they do not appropriately incorporate it into their SPDs. Our results offer new insights into the causes of overconfidence and shed light on potential ways to address this fundamental bias.

1. Introduction

People tend to be too certain of the accuracy of their beliefs. For example, 90% confidence intervals contain the truth as little as 50% of the time—implying that judges are surer of their knowledge than they deserve to be (Lichtenstein et al. 1982). We refer to this as *overprecision*, which is one of several different types of overconfidence (Moore and Healy 2008). Overprecision arises when judges concentrate more probability around their favored answer than accuracy can justify, and correspondingly underestimate the probability that the truth may lie elsewhere. Many studies across disciplines including psychology, decision analysis, and finance have found that subjective probability distributions are typically too narrow and exhibit overprecision (e.g., Russo and Schoemaker 1992; Juslin et al. 1999; Soll and Klayman 2004; Teigen and Jørgensen 2005; Budescu and Du 2007; Glaser and Weber 2007; Ben-David et al. 2013; Jain et al. 2013). At the same time, when there is a distribution of values in a population (e.g., possible outcomes of 500 plays of a gamble), subjective probability distributions (SPDs) are sometimes <u>less</u> concentrated than the true distribution of outcomes (Moore, Carter, and Yang 2015). In this paper, we reconcile this apparent discrepancy. The results offer insight into the psychological processes that lead to overprecision.

The most popular paradigm for studying overprecision asks people to estimate quantities about which they are unsure, such as the weight of a Boeing 787 or the year in which Mozart was born. These types of questions entail *epistemic uncertainty*—uncertainty arising from the judge having only partial information. For example, the judge may know that Mozart was a Classical composer, and that Classical music composition crested in the eighteenth and nineteenth centuries. In contrast, with chance devices such as random number generators, dice, and coin flips, the best anyone can do is to specify a probability distribution of potential outcomes. For instance, the probability of rolling a 7 with a pair of standard dice is 6/36, and the probability of rolling any one of the numbers 6, 7, or 8 is 16/36. This is *aleatory uncertainty*—a representation of an outcome as inherently unpredictable, but with a knowable distribution of probabilities across possible outcomes.

The two types of uncertainty correspond to different reasoning processes, by which people conceptualize an instance either as drawn from a class of events (aleatory) or as a unique and knowable event (epistemic) (Fox and Ülkümen 2011, Ülkümen et al. 2016, Tannenbaum et al. 2017; see also Kahneman and Tversky 1982). For example, most people think of a die roll as an exemplar belonging to a class of possible rolls. Uncertainty is aleatory because, from the perceiver's point of view, any of the possibilities might be realized. In contrast, most people would consider Mozart's birth date to be a unique instance with only one correct answer, as opposed to thinking about Mozart's birth date as a random draw from the distribution of birth dates of Classical composers. The uncertainty here is epistemic because it corresponds to an assessment of one's limited knowledge about Mozart. Different types of events may evoke thoughts of aleatory or epistemic uncertainty, or both, depending on factors such as repeatability (e.g., Mozart can only be born once) and the extent to which outcomes are perceived as resulting from chance versus knowable variables (Nisbett et al. 1983). Epistemic versus aleatory uncertainty captures the distinction between an impression of one's own lack of knowledge and an impression of randomness in the world.

Many judgments involve both epistemic and aleatory uncertainty. For example, how much you should save for retirement depends on how long you will live. People similar to you have a distribution of different lifespans (aleatory uncertainty). At the same time, you may be unsure about what that distribution is or about how your individual features affect it (epistemic uncertainty). Or suppose you are a press photographer aiming to capture the thrill of runners finishing a marathon. Unfortunately, you are less than certain about the expected distribution of finishing times. Your decisions about where and when to position yourself along the course should take into account both the (aleatory) distribution and your (epistemic) uncertainty about that distribution.

In decision analysis, the two types of uncertainty can be assessed separately and then combined according to Bayesian principles (Paté-Cornell 1996). The end result of combining aleatory and epistemic

uncertainty is generally a subjective distribution that is less concentrated than the aleatory distribution is. That's because you are, in effect, considering a distribution of possible distributions. Suppose that you, our press photographer, want to get photos of top competitors finishing the race. You believe, correctly, that there is little aleatory uncertainty. (For example, over the last 50 years, 90% of winning times in the Boston Marathon fell within a 10-minute range (Boston Athletic Association 2023). However, you are not sure where that winning range is (just above 2 hours or just below?). Your subjective probabilities are 30% below 2 hours and 70% above. Your SPD for what might happen is thus much wider than you believe the empirical distribution is. A thorough Bayesian analyst would also incorporate your subjective probabilities for how tightly clustered the win times are and how the distribution is shaped. If done correctly, the judge will be well calibrated, meaning that when they believe there is an X% chance of something, it actually happens X% of the time.

In the research presented here, we examine how people who are not professional decision analysts handle the combination of aleatory and epistemic uncertainties. Most of the prior literature on confidence uses questions that lack aleatory uncertainty because they have a unique correct answer (e.g., "In what year was Mozart born?"). In our studies we assess beliefs about everyday domains in which participants know there is an underlying distribution of outcomes but are less than certain about what the distribution is. Some prior research suggests that laypeople may not combine these types of uncertainty the way decision analysts do. For example, in some studies, people are offered a choice between a purely aleatory gamble, in which the outcome is determined by a random process with known odds, and one that includes a combination of aleatory and epistemic, "second-order" uncertainty, in which the gambler is unsure what the odds are. People prefer the former type even when the two gambles have the same expected utility (Ellsberg, 1961; Baillon and Placido 2019). Other studies show that the subjective chance of success depends on whether the judge has an "outside" view or an "inside" view of a project. The outside view is more aleatory, framing the project as a member of a population of similar projects; the inside view is more

epistemic, framing the project as a unique, risky prospect. The latter framing is more prone to overconfidence (e.g., Kahneman and Lovallo 1993; Epley and Dunning 2006). The present studies build on such findings, developing methods to separate and quantify how people think about aleatory and epistemic uncertainty and how they combine them into subjective probabilities.

2. Two Standards for Subjective Probability Distributions

In our studies, we ask people to estimate a distribution of probabilities for several familiar domains (e.g., commute times, housing prices) in specified cities. For example, we ask questions like, "If we were to randomly choose one person from Philadelphia with a job, what would their average commute time to work be?" A respondent might believe that it is most likely to find a person who commutes between 20 and 30 minutes and less likely to find one who commutes 0 to 10 minutes, or who commutes more than 90 minutes. We are interested in whether SPDs like this are systematically overly concentrated in a few favored ranges (overprecise), too dispersed across many ranges (underprecise), or about right.

In assessing the concentration of SPDs, two different standards apply. When the researcher knows the true distribution of probabilities, it is possible to compare the concentration of the reported subjective distribution to that of the empirical distribution. There is limited research on this topic, but two sets of studies suggest that reported distributions could be more dispersed than the corresponding empirical distributions. Nisbett and Kunda (1985) asked one group of college students to report their own attitudes (e.g., opinion of Ronald Reagan as president) and behaviors (e.g., frequency of going to concerts) and another group to estimate the distribution of one hundred of their peers' answers to those same questions. They found that the standard deviations of the estimated distributions were on average about 10% greater than the empirical ones. Moore et al. (2015) compared subjective and objective distributions for the outcomes of various randomizing devices. For example, they used a Galton board wherein a ball dropped from a slot at the top of the machine bounces over a series of staggered pegs to land in a bin at

the bottom, producing a binomial distribution. They, too, found that subjective distributions for the Galton board as well as other binomial distributions were more dispersed than the objective ones.

Alternatively, we can compare the expressed probabilities of finding a member of a population in a given range (or set of ranges) to the actual probability of finding a member in that range. That is, the judge might assign 25% probability to finding a commuter in the 20-to-30-minute range, when in fact only 15% of commuters fall in that range. To be well calibrated is to have those two probabilities match across a set of questions. Calibration cannot be measured on a single question. If the judge's best guess is right, it will appear that they should have put more probability on that guess (underprecision); if it's wrong, it will appear that they should have put less probability on it (overprecision). Although a well-calibrated judge may appear over- or underprecise on individual questions, on average they assign the right probability (Dawid, 1982). Moore et al. (2015) observed that participants' SPDs fell between the two standards—they were wider than the empirical distribution but not wide enough to be well calibrated.

3. The Present Research

How can the same data show that confidence intervals are too wide by one standard and too narrow by another? Well-calibrated judges do not match the concentrations of their SPDs to the concentrations of the empirical distributions. Their SPDs have to be wider to account for uncertainty about the shapes and locations of the empirical distributions. Consider our sports photographer who is unsure whether the usual finishing time is more or less than 2 hours, or a judge who thinks the 20-to-30 minute range of commutes is most likely, but it might be the 30-40 minute range. These judges must spread subjective probability between the two categories to take both possibilities into account. If judges fail to spread out their SPDs sufficiently to account for such uncertainties, then the concentration of their SPDs can fall between the two standards—less concentrated than the empirical distribution but too concentrated to be well calibrated.

We distinguish between three possible explanations for how people's SPDs can arrive at being between the two standards. First, people may believe that distributions of outcomes in the world are more concentrated than they really are—they underestimate aleatory uncertainty. Although they may have the appropriate intuition that their SPDs should be wider (less concentrated) than they think the empirical distribution is, they widen from an overly narrow base. Second, people may believe that they know the empirical distribution better than they really do. They may disperse their SPDs appropriately given their underestimated degree of epistemic uncertainty. Consistent with this, many authors have argued or implied that overprecision is the direct result of respondents' underestimating their epistemic uncertainty. When Alpert and Raiffa (1982) documented the low hit rates of subjective confidence intervals, they implored their subjects, "For heaven's sake, *Spread Those Extreme Fractiles*! Be honest with yourselves! Admit what you don't know!" (p. 301, emphasis in original). This quote also captures a third explanation. Perhaps people make reasonable estimates of the empirical distribution and of their own epistemic uncertainty. Perhaps they also have the appropriate intuition that epistemic uncertainty should make their SPDs wider. However, they do not know how they should "Spread Those Extreme Fractiles" in light of their uncertainty. To distinguish among these explanations, the next section introduces a novel approach that partitions uncertainty into aleatory and epistemic components. Our new approach affords an opportunity to reconcile discrepant prior findings and gain new insights into the psychology that underlies overprecision in judgment.

4. Measures of concentration and calibration

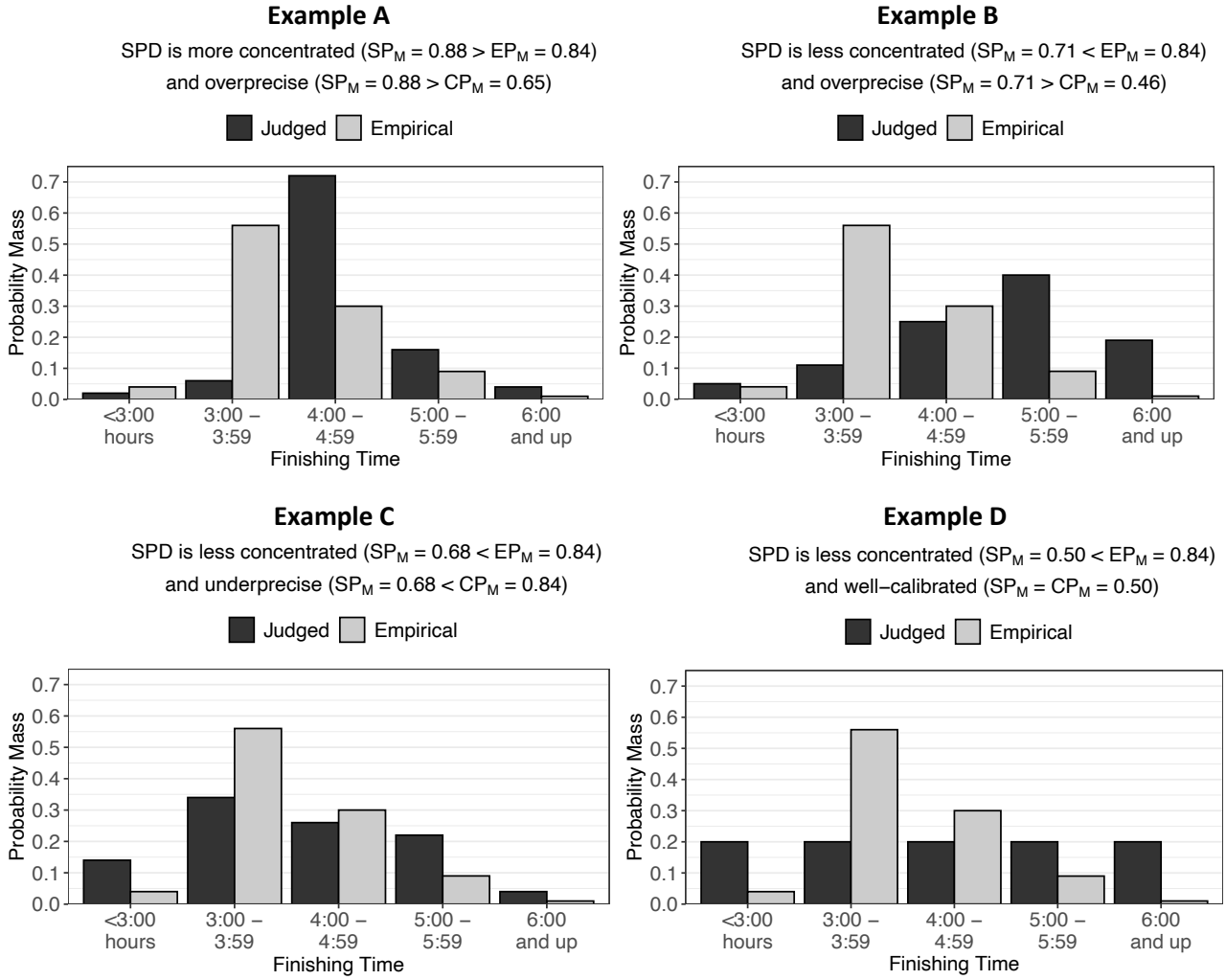
Testing for these three explanations requires us to develop better methods to assess the concentration of SPDs. Research in subjective confidence has asked for X% subjective confidence intervals. However, these provide limited information about the full distribution. An alternative is the histogram elicitation, in which people estimate probabilities for given intervals rather than reporting the interval size corresponding to a fixed probability (Haran et al. 2010; Goldstein and Rothschild 2014). The histogram

elicitation permits researchers to elicit detailed probability distributions by asking participants to assign probabilities to a set of mutually exclusive and exhaustive outcome ranges. Haran et al. (2010) refer to this method as SPIES (for Subjective Probability Interval EstimateS) and show that it produces better-calibrated SPDs compared to earlier methods eliciting a single interval corresponding to a fixed probability. After eliciting distributions using the histogram method, we measure the concentration of the reported distribution.

We then need appropriate measures of concentration and calibration. The standard deviation of a distribution might seem a straightforward measure of concentration. However, when judgments are grouped into categories, as in the histogram method, estimates of standard deviation depend heavily on the shape of the distribution and on assumptions about the distribution of values within categories, especially unbounded end-categories ("greater than __", "less than __"). Therefore, we employ other measures of concentration, which we illustrate with four hypothetical examples. Returning to marathons, Figure 1 shows subjective and empirical probability distributions for four hypothetical judges estimating finishing times across all runners who finish the race. The four examples all assume the same empirical probability distribution, shown in light gray. Each example shows a different subjective probability distribution, shown in dark gray.

Figure 1 shows calculations for three measures of concentration, SP_M, EP_M, and CP_M, which we will explain shortly. First, though, consider a simple measure of concentration, SP_1—the subjective probability the judge assigns to the category they believe is most likely. The judge depicted in Example A believes that the most likely range of finishing times is between 4 and 5 hours, with 72% of runners: $SP_1 = 0.72$. Empirically, that happened only 30% of the time. We call this the *calibrated probability*: $CP_1 = 0.30$. This judge is overprecise on this question: $SP_1 > CP_1$. The difference between the two, $SP_1 – CP_1 = 0.42$, provides a measure of the judge's overprecision ($O_1 = 0.42$) for their top category. We can also compare the concentration of the judge's SPD with the concentration of the true, empirical distribution, which we call

Figure 1. Subjective and empirical probability distributions for four hypothetical judges estimating marathon finishing times.



EP_1. The judge believes that 72% of finishing times fall into the most likely category whereas empirically only 56% of finishing times fall in the category that is really most likely (3 to 4 hours). The judge's SPD is more concentrated than the empirical distribution is: $SP_1 = 0.72 > EP_1 = 0.56$.

SP_1, CP_1, and EP_1 are useful, but we can obtain a more complete picture by using cumulative measures that take more categories into account. We could look at the judge's top two categories rather than just one, which together comprise times between 4 and 6 hours. Our judge assigned a probability of 0.88 to that range: $SP_2 = 0.72 + 0.16$. Looking at the empirical distribution, we can see that the calibrated probability for the range of 4 to 6 hours is only 0.39, $CP_2 = 0.30 + 0.09$, and the resulting overprecision is

O_2 = 0.49. In the empirical distribution, the top two categories are actually 3 to 4 hours and 4 to 5 hours. They include 56% + 30% of finishing times, so EP_2 = 0.86. We can similarly compute SP_3, CP_3, EP_3, and O_3 looking at the top three categories and SP_4, CP_4, EP_4, and O_4 using the top four.

Example A illustrates that conclusions can differ depending on which level of cumulation one uses. With only the single, top category, the judge's SPD looks much more concentrated than the empirical distribution is ($SP_1 = 0.72 > EP_1 = 0.56$). If we look at the top two categories, though, the two concentrations are about the same ($SP_2 = 0.88 \approx EP_2 = 0.86$). Which is the appropriate number of categories for a measure of calibration? There is no right answer to this question, so we average across them to get *mean cumulative concentration* (MCC) measures: SP_M for the judge's subjective probabilities, CP_M for the calibrated probabilities, and EP_M for the empirical probabilities. In Example A, $SP_M = (SP_1 + SP_2 + SP_3 + SP_4)/4 = 0.88$. This captures the extent to which the judge concentrates probability mass in a few categories or spreads probability mass out across all categories. If a judge assigns all of the mass to one category, then SP = 1 at every level of cumulation, and $SP_M = 1$. At the other extreme, if a judge thinks that all categories are equally likely, as in Example D, then $SP_M = 0.5$ (= (0.2 + 0.4 + 0.6 + 0.8)/4). We calculate CP_M and EP_M similarly, using the CP and EP measures, respectively. In Example A, the judge's SPD is more concentrated than the empirical distribution ($SP_M > EP_M$), and is overprecise, that is, more concentrated than it should be for good calibration ($SP_M > CP_M$). The average amount of overprecision in Example A can be calculated as $O_M = SP_M - CP_M = 0.23$.

The four examples in Figure 1 illustrate that different combinations of (over)precision and relative concentration are possible. Example B illustrates the two-standards paradox described earlier. This judge's SPD is less concentrated than the empirical distribution: $SP_M < EP_M$. This accords with Bayesian norms, yet Judge B is still overprecise: $SP_M > CP_M$. To be well calibrated, Judge B needed to widen their SPD more, assigning less probability to their favored categories and more to less favored categories. In Example C, the judge has overdone it. They have spread out their subjective distribution so much that

they are underprecise. In this instance, the judge would have done better to assign more probability mass to their favored categories. Example D shows the extreme case in which the judge expresses no idea which categories are more or less likely. SP_M is at the minimum possible value of 0.5, so this judge's distribution is, of course, less concentrated than the empirical distribution. Interestingly, a uniform SPD always yields perfect calibration. When a judge indicates that they have no idea which categories are more likely, their probabilities are appropriately spread out.[1]

Although mean cumulative concentration is a novel method for measuring and evaluating the concentration of probability mass, it is equivalent to a measure frequently used in economics to measure concentration of wealth: the Gini coefficient (for further details, see §S2.1 of the supplement). The MCC measures have significant advantages. Other measures of overprecision differ depending on the confidence level being assessed (e.g., 90% vs. 50% intervals; Teigen and Jørgensen 2005). MCC measures provide a global picture of overprecision by averaging across different, successive sub-intervals, and can be used across a variety of variable types and elicitation formats: binary, continuous, multiple choice, or ordinal. One caveat, however, is that the measures will vary to some degree depending on the number of categories dividing up a continuous scale, and on where the "greater than ___" and "less than ___" end-categories begin.[2] Our studies mitigate this concern by comparing coefficients from comparable partitions and by using pre-existing, externally determined partitions when possible.

---

[1] In §S2.1 of the supplement, we offer proofs of the two results that we have informally described here. Proposition 1 shows that a perfectly flat distribution is well-calibrated. In practice, whenever there are ties in a participant's ranking of the likelihood of categories (as revealed by their SPD), we simulate 1,000 rank orderings and calculate CP_1 and CP_M by averaging over these. This approximates perfect calibration for flat distributions and makes our results stable when the same data are reanalyzed. Proposition 2 specifies a requirement for good calibration. Formally, a Bayesian judge would anticipate all possible empirical distributions, each with a different EP_M, and calculate an expected value over these. For a well-calibrated judge, SP_M will be less than this expected EP_M.

[2] It is also the case that a small change to an SPD, one that changes the rank-ordering of bin-likelihoods, can potentially have a large effect on CP_M. For this reason, calibration should always be evaluated across a large set of questions so that performance on individual questions averages out, as is generally true of calibration studies.

5. Hypotheses and Overview of Experiments

In our experiments, we compare the concentration of the judge's distribution (SP_M) with that of the empirical distribution (EP_M). To be well-calibrated, judges should spread out their SPDs so that they are less concentrated than they expect the empirical distribution to be. That is, SP_M should be smaller than EP_M. If judges do not do that, or do so insufficiently, their judgments will tend to be overprecise. If we find that SP_M is actually larger than EP_M, this would suggest that overprecision is at least partially due to people underestimating the variability in real-world outcomes, such as believing that steep gains and losses on a stock are less likely than they really are. Even if SP_M is less than EP_M, though, judges might still insufficiently account for epistemic uncertainty, and this also can lead to overprecision.

Our first two experiments seek to unpack the relationships among concentration, epistemic uncertainty, and overprecision. As we have noted, previous research has found that people's subjective probabilities are less concentrated than the empirical distribution, but more concentrated than necessary for good calibration. In demonstrating these effects, Moore et al. (2015) asked participants to assess SPDs for stochastic devices such as the Galton board. Experiments 1 and 2 test the generality of this finding by comparing the concentrations of participants' SPDs with those of the empirical distributions across a number of different domains of knowledge such as daily commute times or high temperatures in different cities. A second goal of these initial experiments was to examine whether overprecision could be understand as a directionally correct, but insufficient, response to epistemic uncertainty about the empirical distribution. To test for this, we manipulated knowledge by providing participants with either the median (Experiment 1) or the mode (Experiment 2) of the distribution. If participants are sensitive to epistemic uncertainty, their SPDs should be less concentrated the less they know about the distribution. They might still be overprecise, though, if they do not lower concentration sufficiently to account for their lack of knowledge.

Based on the above reasoning and prior literature, we went into Experiments 1 and 2 with three hypotheses in mind:

(H1) $SP_1 < EP_1$ and $SP_M < EP_M$. As observed in the Nisbett and Kunda (1985) and Moore et al. (2015) studies, SPDs are less concentrated than the true distribution.

(H2) SP_1 and SP_M are lower when epistemic uncertainty is greater. The less judges know about the empirical distribution (e.g., not knowing the median or mode), the more they should spread out their SPDs.

(H3) $O_1 = SP_1 – CP_1 > 0$ and $O_M = SP_M - CP_M > 0$. Even if participants widen their distributions where there is epistemic uncertainty, that they will do so insufficiently. Therefore, SPDs are more concentrated than they would need to be for good calibration.

Although we find substantial overprecision (H3) in Experiments 1 and 2, we find little support for H1 or H2. To further explore the apparent insensitivity to epistemic uncertainty observed in those studies, Experiments 3 and 4 disentangle beliefs about the distribution from uncertainty about those beliefs by having participants directly estimate the concentration of the empirical distribution. Those experiments show that people do incorporate their epistemic uncertainty in their SPDs, although not in a way that facilitates good calibration. It seems that people confound being certain about the empirical distribution with being certain about where in the distribution an item will fall. This leads to excessive concentration and overprecision. We conclude that multiple factors contribute to overprecision, making it particularly challenging to de-bias.

Pre-registrations, experimental materials, and data for all four experiments are available at https://osf.io/dt7cq. We report results for all conditions and dependent variables. In some instances, alternative analyses are provided in the electronic supplement.

6. Experiment 1

In this experiment, we manipulated the provision of information. Some participants learned the median of the empirical distribution and others did not. Information about the median does not eliminate all epistemic uncertainty: Participants will still be uncertain about how probability is distributed among the categories in the empirical distribution. However, knowing the median reduces epistemic uncertainty for nearly all participants, to varying degrees, because it reduces uncertainty about where the empirical distribution is centered along the scale. Most participants who correctly guess the median value will not be 100% certain about it; giving them the median provides additional evidence that they are right, allowing them to be more confident in their guesses. A key test is whether participants who lack the median spread out their SPDs to account for epistemic uncertainty. We anticipated that people insufficiently do so, and that this contributes to overprecision.

In this experiment and the subsequent ones, we asked participants about variables sampled from a particular city, selected at random for each domain of knowledge from a list of 40 large U.S. cities. For example, one participant might be asked about the income of a randomly drawn household in Seattle, Washington while another might be asked about the income of a randomly drawn household in Atlanta, Georgia. We selected cities in this way to reduce concerns that observed overprecision could be attributable to the over-representation of surprising, "contrary" questions—ones for which usually-valid information or intuition points to an incorrect answer (Klayman et al. 1999).

6.1. Method

6.1.1. Participants. Based on a pre-registered power analysis, we aimed to recruit an MTurk sample size of 600, split evenly between two conditions. Of the 973 people who began the online survey, 594 successfully completed the training and finished the study. Although some participants dropped out after being randomly assigned to a condition, they were roughly evenly divided between conditions. Age and

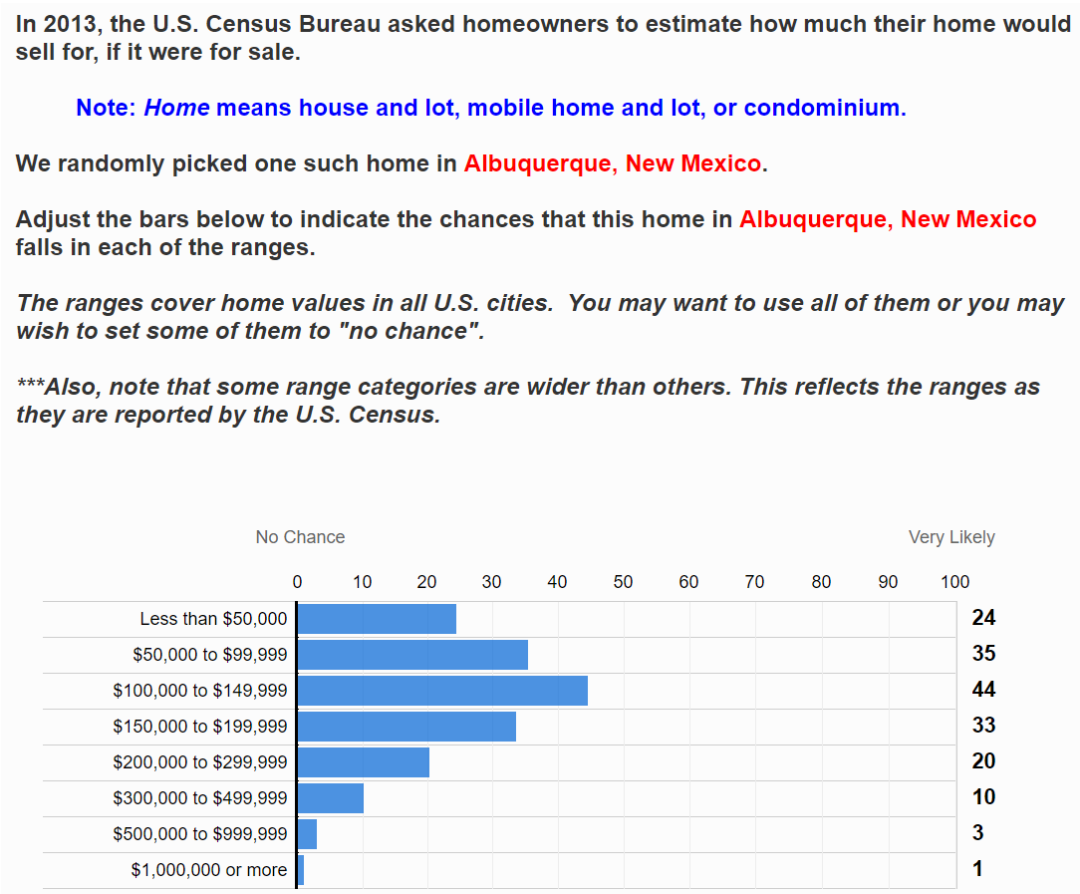
gender were not collected in Experiment 1. Participants received a base payment of $0.50 and an average bonus of $0.55 for accuracy.

6.1.2. Materials. Figure 2 shows an example question. Participants reported their subjective likelihoods by adjusting the slider bars, which participants were explicitly told did not need to add up to 100. Rather, they were instructed to adjust the bars to indicate the relative chance of an observation being in that category (e.g., a bar three times as long means that it is three times as likely). We normalized reported distributions by dividing each category's assigned likelihood by the total across all categories. For each domain, the spectrum of possible answers was divided into a modest number of response categories, as with the eight categories shown in Figure 2. Four of the domains use the 7 to 12 categories by which the U.S. Census Bureau reported data from the American Community Survey; for the fifth domain, average daily high temperatures, we defined categories in increments of 10 from 0 to 100 Fahrenheit, with end-ranges of "less than or equal to 0" and "greater than 100."

6.1.3. Procedure. A practice item taught participants how to use and interpret the slider bars. They then took a 3-item quiz to make sure that they understood that longer bars represented greater chances, that the relative lengths of bars represented relative chances, and that the bars did not need to sum to 100. Participants had two tries to correctly answer each question and had to answer all three questions correctly in order to be included in the analysis. Those who were provided with medians received a brief explanation of the median and a quiz question to test their understanding. To ensure that we could generalize the sample in each condition to the same population, we did not exclude anyone based on this question. After exclusions, 310 participants were provided with medians and 284 were not.

Next, we told participants to "set the bars to reflect your true beliefs about the relative chances that a random observation will fall in each category. The more accurate your responses, the higher your

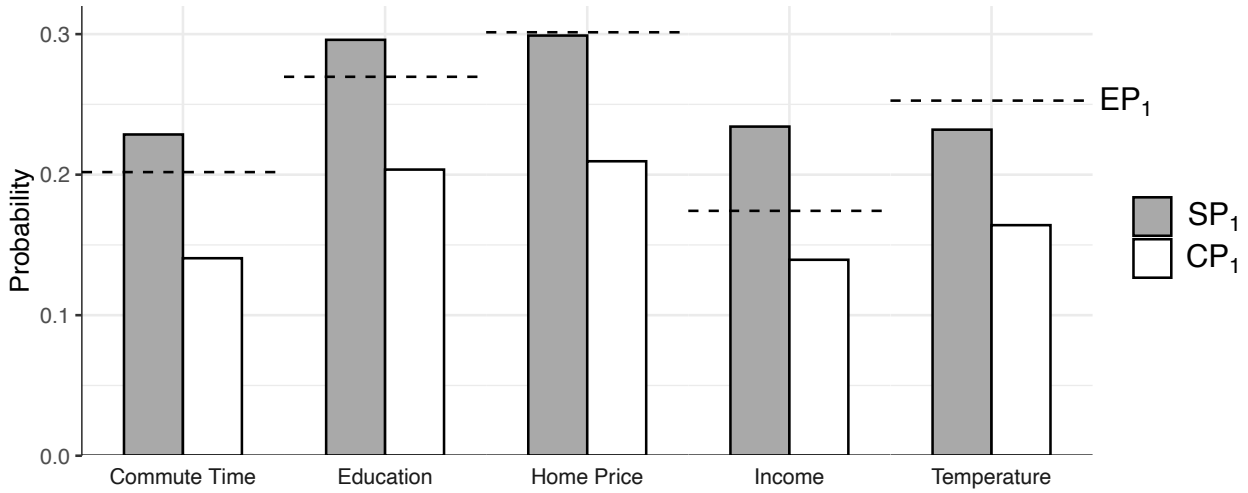
Figure 2. Example question with responses from Experiment 1.



bonus will be." We did not provide them with the details of the payoff formula.[3] After the instructions and comprehension questions, participants saw one question from each of the five domains, presented in random order, with a new city randomly selected for each question. At the conclusion of the study, participants were debriefed and informed that we would deliver base payments within 24 hours and bonuses within one week.

[3] Bonus payments were calculated for each question using an incentive-compatible extension of the Brier (1950) score. For each of the *n* total categories *c* available for the question, we calculated the quadratic score the judge would receive if a randomly-chosen instance were to fall in that category: $B_c = \sum_{i=1}^{n}(\boldsymbol{I}_i^c - \hat{p}_i)^2$, where $\hat{p}_i$ is the probability the participant assigned to category *i* and the indicator $\boldsymbol{I}_i^c$ equals 1 when *i* = *c* and 0 otherwise. We then multiplied the quadratic score for each category *c* by the empirical probability p_c that a randomly chosen member of the population would, in fact, fall in that category and summed those products to arrive at an average score for that question: $EB = \sum_{c=1}^{n} p_c B_c$. The bonus earned was $20(1 - EB)$ cents for each question.

Figure 3. Most-likely-category probabilities in Experiment 1.



Note. The gray bars represent SP_1 (the highest probability assigned to a category) and the white bars represent CP_1 (the empirical probability of that same category). The dashed lines indicate EP_1, the probability of the most likely category in the empirical distribution.

6.2. Results

We present descriptive results for two measures of concentration—the probability assigned to the most likely category and the overall concentration of the distribution. Statistical analyses for the two were essentially identical, so we report those for overall concentration here and those for the most likely category in the supplement.

Descriptive results for the most likely category are shown in Figure 3, collapsing over the two levels of Information. Providing the median had very little effect on judgments. Contrary to H1, participants' SPDs were significantly more concentrated than the empirical distribution ($SP_1 > EP_1$) for three of the domains and were significantly less concentrated only for temperatures. However, participants were overconfident for all five domains, meaning that participants assigned more probability to their favored category than the empirical distribution did to that same category ($SP_1 > CP_1$). This was even true in the temperature domain where participants' judgments were less concentrated than the empirical distribution.

Table 1. Mean cumulative concentrations and overprecision in Experiment 1.

Domain	**EP_M**		**SP_M**		**CP_M**		**O_M**	
Medians Provided	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Commutes	.655	.656	.683	.693	.544	.581	.140	.112
Education	.673	.672	.706	.697	.580	.587	.125	.111
Home Prices	.738	.744	.736	.732	.622	.670	.114	.062
Incomes	.633	.633	.718	.708	.581	.587	.137	.121
Temperatures	.800	.801	.757	.760	.689	.731	.068	.029
Mean	**.700**	**.701**	**.720**	**.718**	**.603**	**.631**	**.117**	**.087**
Overall Mean	**.700**		**.719**		**.617**		**.103**	

The means for overall concentration are shown in Table 1. We conducted mixed-model ANOVAs with Information (median provided vs. not) as a between-subjects variable and Domain as a within-subjects variable. To test H1, we use the difference score SP_M – EP_M as the dependent variable. Contrary to H1, participants' SPDs were more concentrated than the empirical distribution, $F_{1, 592} = 82.5$, $p < .001$, $\eta_p^2 = .122$. There was also a main effect of Domain, $F_{3.63, 2146} = 324.1$, $p < .001$, $\eta_p^2 = .354$, reflecting the fact that SPDs were more concentrated than the empirical distribution in some domains and less concentrated in others. There was no effect of Information, $F_{1, 592} = 0.75$, $p = .387$. SPDs were more concentrated than the empirical distribution regardless of whether participants learned the median.

Next, SP_M showed a main effect of Domain ($F_{3.65, 2163} = 160.9$, $p < .001$, $\eta_p^2 = .214$) but no main effect of Information. We did find an unexpected Domain x Information interaction ($F_{3.65, 2163} = 2.62$, $p = .011$, $\eta_p^2 = .006$), which captures small differences between domains in the different information conditions. Regardless, provision of the median had little effect on the concentration of participants' SPDs, and consequently H2 was not supported.

Participants were overprecise overall: In support of H3, mean O_M exceeded zero, $F_{1, 592} = 1{,}949$, $p < .001$, $\eta_p^2 = .767$. Overprecision was less when participants were provided with the median, $F_{1, 592} = 41.2$, $p < .001$, $\eta_p^2 = .065$. As shown in Table 1, this was a consequence of judges' estimates being closer to the

truth when they were given the median (i.e., greater CP_M) without being more concentrated (i.e., no difference in SP_M). A main effect of domain ($F_{3.18,\ 1882}$ = 79.9, p < .001, η_p^2 = .119) and a small Domain x Information interaction ($F_{3.18,\ 1882}$ = 4.47, p = .003, η_p^2 = .007) indicate that overprecision differed by domain.

6.3 Discussion

In accord with prior research, Experiment 1 shows that people are consistently overprecise—their SPDs concentrate too much probability in too little of the spectrum. To be well calibrated, SPDs must be wider than the underlying empirical distribution, because they must reflect both the variability in the empirical distribution (aleatory uncertainty) and the likelihood of errors in estimating what that distribution is (epistemic uncertainty). Our results show that SPDs are, on the contrary, slightly *narrower* than their corresponding empirical distributions. More importantly, there are large differences among domains, suggesting that it is better to think of this comparison as a characteristic of a specific domain of judgment rather than as a pervasive tendency.

In the introduction, we posited three explanations for why SPDs are not wide enough for good calibration: (a) On average, people might believe empirical distributions to be more concentrated than they really are; (b) people might be too epistemically certain–that is, they think they know more about the empirical distribution than they do; and (c) although epistemic uncertainty would demand that people widen their SPDs, people fail to do so. The results imply that (a) is not a sufficient explanation. Judges who overestimate empirical concentrations would tend to be overprecise, but they should still have wider SPDs when epistemic uncertainty is greater. Explanation (b) is plausible if participants who were not told the median failed to recognize their lack of knowledge. Explanation (c) is plausible if participants not told the median recognized their lack of knowledge but nonetheless did not know what to do about it. Experiment 2 tests whether people widen their SPDs when lack of knowledge is more salient.

7. Experiment 2

In this experiment, we sought to increase the salience of epistemic uncertainty by having participants experience either the addition or the deletion of relevant information about the distribution. All participants estimated probabilities for a randomly selected exemplar drawn from each of six domains, presented in two sets of three, in a procedure similar to that of Experiment 1. One group was told the modal category for each domain (*mode–mode)*; another group was not given that information (*no mode–no mode)*. A third group received the mode for the first three exemplars, but not for the last three (*mode–no mode)*; a fourth group encountered the reverse pattern (*no mode–mode)*. We predicted that losing or gaining information would make participants more aware of their state of knowledge following the change. Thus, if participants respond appropriately to epistemic uncertainty, but need to be prompted to think of it, an obvious change in available information should cue them to reduce the concentration of their subjective distributions when information is removed and to increase the concentration when information is added.

We modified the materials of Experiment 1 in two ways intended to increase participants' awareness of epistemic uncertainty. Instead of providing the median we provided the mode as additional information. Arguably, the mode is more useful because it tells participants the most likely category, whereas the median provides only a strong hint about which one it might be. Second, we asked participants for their "confidence" for each category as opposed to its "chances." Because people associate the term "confidence" with epistemic uncertainty (Tannenbaum et al. 2017), specifically asking for this might prime participants to account for it.

7.1 Method

7.1.1. Participants. In this study, participants were assigned to a condition only if they successfully passed the training. Of the 1,166 MTurk participants who began the online survey, 430 failed to successfully complete the training, either by dropping out (79) or by failing to meet the criterion for

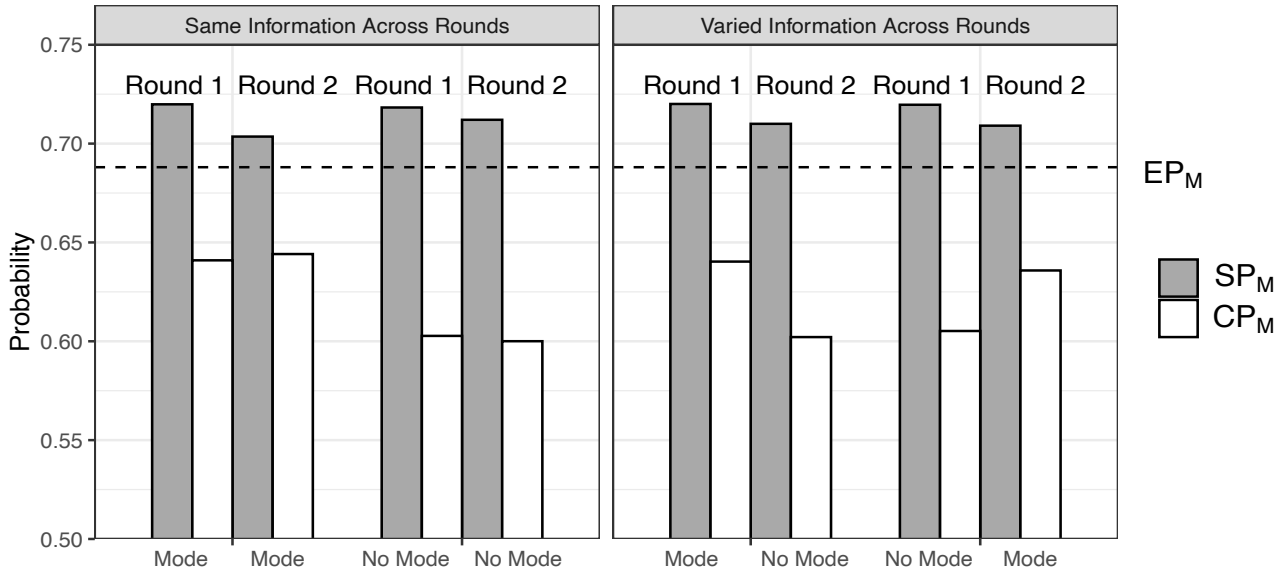
passing (351). Another 23 passed the training but failed to properly complete the study. In accord with our pre-registered design, analyses included the first two qualified participants for each of the 84 unique stimulus sets in each of the 4 conditions. In the final sample of 672, the mean age was 38.3 (*SD* = 12.5), and 54.2% were female.

7.1.2. Design. The experiment had a 2 (Round) x 4 (Information) mixed factorial design. Round was a within-subjects factor, with three domains per round. We systematically varied the order of the six domains using a Latin square, yielding six different orders. Each order was duplicated seven times using a different set of six randomly selected cities to use with the six domains. We then duplicated the resulting 42 sets of questions, interchanging the rounds (i.e., questions 1-3 became 4-6, and vice versa), creating 84 sets of 6 questions each. Information condition varied between subjects: Participants were told the mode for either all or none of the three Round 1 questions and all or none of the three Round 2 questions.

7.1.3. Materials and Procedure. We drew questions from the five domains used in Experiment 1, and added a sixth domain, the age of a randomly chosen individual in the selected city. Participants underwent similar training as in Experiment 1. They read that they could maximize their bonus by setting the bars to reflect their true degree of confidence. We used the same payment scheme as in Experiment 1. Participants received a base payment of $0.75 and an average bonus of $0.67 for accuracy.

At the beginning of Round 1, participants who received the mode read, “For the first three items we will provide you with some helpful information. We will tell you the most common category for the given city. Please click below to go to the first item.” Between Rounds 1 and 2 they learned either “For the next three items we will continue to tell you…” (mode–mode condition) or “For the next three items we will no longer provide you with the additional information. We will not tell you…” (mode–no mode condition). For participants who were not given the mode, Round 1 began with just the instruction, “Please click below to go to the first item.” Those in the no mode–no mode condition received no additional instruction at the start of Round 2. For those in the no mode–mode condition, Round 2 had the

Figure 4. Concentration and calibration in Experiment 2.



Note. The left panel shows the two information conditions in which participants saw the same information throughout the study. The right panel shows the two conditions in which the provision of information changed between rounds.

same instructions given at the beginning of the mode–mode condition, "For the next three items we will provide you with some helpful information…."

7.2 Results

As in Experiment 1, we analyzed concentration by looking at the most likely category and at the entire distribution. Results are similar, so we focus here on the latter and report the former in the supplement. For each concentration measure, we averaged across the three domains within each round. This allowed us to analyze the data as a 4 (Information condition) X 2 (Round) mixed model ANOVA.

Because the design was perfectly balanced, the average EP_M was 0.688 in all conditions and rounds. Overall, participants' distributions were more concentrated than the empirical ones: The mean difference between SP_M and EP_M was 0.026, $F_{1, 668} = 291$, $p < .001$, $\eta_p^2 = .303$. This is clear in Figure 4, which shows that all the gray bars (SP_M) are above the horizontal line (EP_M). As in Experiment 1, domains varied in whether participants were more or less concentrated than the empirical distribution (see supplement). Overall, however, H1, that SPDs are less concentrated than the true distribution, was not supported.

Overprecision ($O_M = SP_M - CP_M$) was pervasive, with a mean level of 0.093 that significantly differed from zero, $F_{1, 668}$ = 2,548, $p < .001$, η_p^2 = .792. The critical result is a Round x Information interaction ($F_{3, 668}$ = 22.5, $p < .001$, η_p^2 = .092), such that participants were more overprecise whenever they lacked the mode. As shown in Figure 4, this happened because CP_M (shown by white bars in Figure 4) was higher whenever the mode was given, indicating that participants' SPDs more closely matched the empirical distribution. However, this is not true of SP_M (gray bars), indicating that the concentration of SPDs varied little with changes in epistemic uncertainty.[4]

7.3 Discussion

Overall, the results are consistent with those of Experiment 1. Participants' distributions were slightly more concentrated than the empirical distributions, and much more concentrated than needed to be well calibrated. Participants were better calibrated when provided with the mode of the distribution. However, despite participants' awareness of their knowledge state, they did not translate this into the concentration of their subjective distributions. Even a strong hint in the form of previously available information being taken away did not have a measurable impact on the concentration of SPDs.

8. Experiment 3

Experiment 2 showed that even when epistemic uncertainty was made salient by adding or removing information, participants did not, on average, widen their SPDs. This could be because they did not understand that the appropriate response to epistemic uncertainty is always to widen their SPDs. Alternatively, participants may not have understood that the SPD is, in principle, different from one's best guess of the distribution.

[4] There was also a main effect of Round that we did not predict. Participants' SPDs were on average less concentrated as measured by SP_M in Round 2 than in Round 1, $F_{1, 668}$ = 29.8, $p < .001$, η_p^2 = .043. This led to slightly less overprecision in Round 2, $F_{1, 668}$ = 8.6, p = .003, η_p^2 = .013.

In this experiment, we sought to make the distinction between the SPD and one's best guess about the shape of the distribution as clear as we could, and we elicited estimates for both. Rather than elicit the entire belief distribution, we asked participants to focus on the single most likely range category for a randomly selected exemplar (e.g., the most likely category for the commute time of a randomly selected working adult in Austin, Texas). We then manipulated the type of estimate that participants were asked to make. We asked one group of participants (the specific-bin condition) to choose the category that they believed was most likely for the exemplar, and then to estimate the chances that the exemplar would be in that specific category. As in our previous studies, this estimate (SP_1) should take into account both aleatory uncertainty (what proportion of the population is in the likeliest category) and epistemic uncertainty (the possibility that some other category is actually more likely). We asked another group (the generic-bin condition) to estimate the percentage of members of the population that are found in the most common category, *whichever category that happens to be*. We also asked about half of the participants to tell us how confident they were in their belief about which category was empirically most likely. This was meant to cue those participants into epistemic uncertainty, which we thought might lead to lower probability estimates.

The key difference between the specific-bin and generic-bin conditions is the source of epistemic uncertainty that judges must consider in order to be well calibrated. Take, for instance, a judge in the specific-bin condition whose best guess is that the most common category for commute times is 20-30 minutes, and that 40% of workers have commutes that fall in that range. As in our previous studies, their subjective probability of encountering a commuter in the 20-30 minutes category (SP_1) should take into account the possibility that the most common category is actually shorter or longer than 20-30 minutes. Accordingly, our judge should be sensitive to how sure they are about which category is empirically most likely. The more uncertain the judge is about whether 20-30 minutes is really the most common, the more they should moderate their subjective probability of finding a commuter in that range. Now consider a

judge in the generic-bin condition whose best guess is that 40% of commuters fall into the most common category, *whichever category that happens to be*. This judge has some epistemic uncertainty about whether 40% is right but does not have to worry about whether they know which category is empirically most likely. Thus, subjective probabilities in the generic-bin condition should be much less sensitive to uncertainty about which is the peak category.

8.1 Method

8.1.1. Participants. Our pre-registered research plan called for a sample size of 960 participants from the ROI Rocket – ClearVoice online research panel. To ensure that participants were comfortable working with numerical information, they had to pass a 5-question math quiz to continue to the main study (see supplement). Only those who answered at least four questions correctly were randomly assigned to a condition; those who failed were shown their responses along with the correct answers and informed that they could not continue. About 60% of potential participants (1,682 out of 2,720) passed the quiz and were randomly assigned to a condition. An additional 131 participants in the generic-bin condition and 140 in the specific-bin condition dropped out during the course of the study. Prior to analysis we randomly dropped participants in groups where quotas were surpassed until the cell size was as pre-registered. In addition to a standard base payment of $0.50 from the survey company, participants also received an average bonus of $0.87 for accuracy. Participants' average age was 49.3 (*SD* = 11.8), and 65% were female.

8.1.2. Design. Each participant responded to six questions from the same set of domains and cities as in Experiments 1 and 2. Participants were evenly divided across the four between-subjects conditions in a 2 (Judgment Type: specific vs. generic bin) x 2 (Prompt: present vs. absent) design. Of those who received a prompt, half reported how confident they were about the most common category before they provided each estimate and half were asked after. (As there was no effect of timing, we will report analyses that collapse across this variable.) We created sets of questions using a Latin square for the order

of domains, and six different cities were randomly chosen for each order. We repeated this process 20 times using the same original square of domain orders but pairing with new groups of cities. This resulted in 120 unique sets of six questions to present to participants. The total collection of 720 questions (120 sets times 6 questions per set) featured each of the 40 candidate cities in 18 of these sets. No set included the same city more than once.

8.1.3. Materials and procedures. After passing the math quiz, participants received instructions appropriate to their condition, along with several questions to make sure that they understood the basic concepts of subjective probability and understood all the elements of the procedure. Participants then saw the correct answers to these comprehension questions alongside their own. (They were included in the analysis regardless of how well they did on these.) Participants then read that they should try to be as accurate as possible, and that they could earn up to an additional $0.20 cents for each question based on their accuracy.[5]

The estimate questions were like those used in Experiment 2, except that we reorganized the ranges of values in five of the six domains so that each question would have five categories with the three middle categories being of equal width. For example, the categories for commute times were < 15, 15-29, 30-44, 45-59, and ≥ 60 minutes. These modifications served to simplify the task and to make results for different questions more comparable. For the education domain, which is categorical, we simply reduced the number of categories to five (e.g., two of the original categories were combined into "Did not complete high school").

In the generic-bin condition, the estimate questions took this form: "In Atlanta, Georgia, what is the percentage of adult workers in the *most common* category of commuting times, <u>whichever category</u>

[5] As in Experiments 1 and 2, bonus payments were calculated using an incentive-compatible extension of the Brier score. For each item, participants provided a probability judgment $\hat{p}$ (for either *the category they chose*, or *the most likely category, whichever category that happens to be*, in the probability and concentration conditions, respectively). We used the true probability p for that particular question to calculate an expected Brier score according to $EB = p(1-\hat{p})^2 + (1-p)\hat{p}^2$. The bonus earned was $20(1-EB)$ cents for each of the six items.

that happens to be?" (Emphases were included.) Participants responded by selecting one of 21 radio buttons labeled 0% to 100% in increments of 5%. Those who received a prompt for this question were asked this either before or after each question: "How confident are you that you could correctly choose the most common category of commuting times in Atlanta, Georgia." They responded by selecting one of five confidence levels, ranging from "Not at all confident" to "Extremely confident."

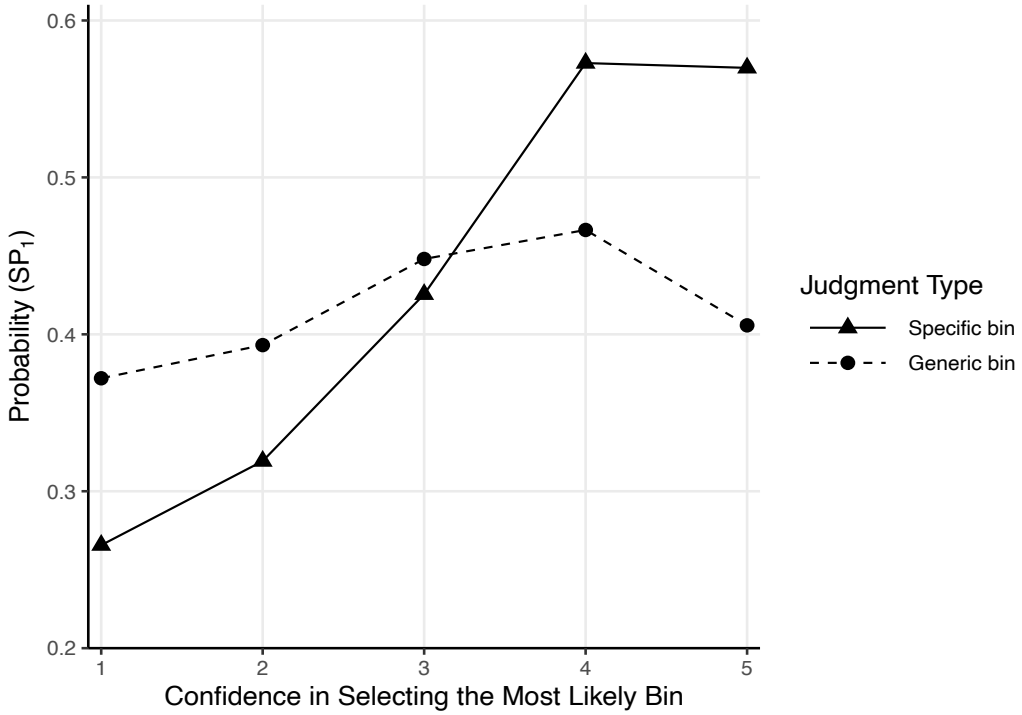
In the specific-bin condition, estimates were elicited using a two-part question, such as, "In your judgment, which of the five commuting time categories is the *most common* one in Atlanta, Georgia?", followed by, "We're going to select a working adult in Atlanta, Georgia at random. What are the chances that this person was in the commuting time category that you chose?" They responded using the same 0%-100% scale as in the modal category condition. Participants who received a prompt were also asked (either before or after the probability question), "How confident are you that you correctly chose the *most common* commuting time category in Atlanta, Georgia?" They responded on the same 5-point confidence scale described above.

8.2 Results

For these analyses we work again with SP_1, CP_1, and EP_1, which correspond, respectively, to the subjective probability assigned to the category judged as most likely, the empirical probability associated with that category, and the empirical probability associated with the actual most likely category. Note that for that for the generic-bin condition it must be the case that $CP_1 = EP_1$ because participants are estimating the category that is, in fact, most likely (whatever that may be). For the specific-bin condition, $CP_1 = EP_1$ when participants correctly guess the most common category, and otherwise $CP_1 < EP_1$.

We analyzed these data with a 2 (Judgment Type) x 2 (Prompt) x 6 (Domain) mixed ANOVA design. Overall, participants' judgments were about as concentrated as the empirical distributions: $SP_1 – EP_1$ did not differ significantly from zero ($F_{1, 956} = 1.53$, $p = .22$), nor did they differ between Judgment Type and

Figure 5. Most-likely-category probabilities in Experiment 3, by type of event and confidence level.



Note. *Confidence* = confidence in identifying the most common category, *Probability* = estimates of the probability of a randomly chosen item falling into the category believed most likely (specific bin), or falling in the most likely category, whichever category that is (generic bin).

Prompt conditions. However, as in the first two experiments, there were large domain differences, $F_{4.4, 4203} = 188$, $p < .001$, $\eta_p^2 = 0.164$. Also, participants were more overprecise ($SP_1 - CP_1$ was larger) when judging a specific versus a generic bin ($M = 0.129$ vs. 0.014), $F_{1, 956} = 99.8$, $p < .001$, $\eta_p^2 = 0.095$.

Our primary interest in this study was whether SP_1 varied across conditions. If people are sensitive to epistemic uncertainty, they should report lower probabilities when thinking of specific as opposed to generic bins. Especially in the prompt condition, we anticipated that people might recognize and account for epistemic uncertainty to some degree. On this first look, we found no hint of effects for either Judgment Type ($F_{1, 956} = 1.60$, $p = .21$, $\eta_p^2 = 0.002$) or Prompt ($F_{1, 956} = 0.003$, $p = .96$, $\eta_p^2 = 0.000$). It appears that participants failed to account for epistemic uncertainty, even when prompted to do so.

However, when we take into account participants' confidence in correctly identifying the most common category, a different pattern emerges (see Figure 5). For this regression analysis, which was not

preregistered, we could only use the data from participants who were prompted to provide their confidence in their best guess about the most common category.

We regressed estimates on Judgment Type (0 = generic bin, 1 = specific bin), Domain (effects coded) and mean-centered Confidence (M = 2.70, SD = 1.02), along with the Judgment Type x Confidence interaction. Random intercepts were included for participants to account for correlated observations. The interaction was significant (t_{479} = 4.66, p < .001), reflecting a steeper slope for estimating specific as opposed to generic bins. Participants were sensitive to epistemic uncertainty in the specific-bin condition, where it was appropriate to be so. They were also significantly less sensitive in the generic-bin condition, where reducing concentration in the face of epistemic uncertainty is not mandated. Nevertheless, as reported earlier, participants in the specific bin condition were overprecise. These results are consistent with a directionally correct but insufficient response to epistemic uncertainty. However, a spotlight analysis at high confidence (1.5 SD above the mean, at 4.24) revealed a hitch. Here, participants estimated a *higher* probability for a specific bin than for the comparable generic bin, b = .096, t_{479} = 2.97, p = .003. This result is not consistent with just insufficient adjustment. It contradicts the Bayesian prescription that one must report a lower subjective probability unless one is absolutely certain about which category is most common, in which case they are equal.

8.3. Replication

The results shown in Figure 5 were a surprise, so we conducted an exact replication of the experiment. The results were the same—probabilities for specific bins rose more steeply with confidence than did those for generic bins. Moreover, probabilities were lower for specific bins than for generic bins when confidence was low, and higher for specific bins than for generic bins when confidence was high (see §S1.3 of the supplement for details).

8.4. Discussion

Subjective probability judgments should take into account both the aleatory distribution of possibilities in the population and one's epistemic uncertainty about what that distribution is. In Experiments 1 and 2, we found little evidence that people follow this normative rule. Experiment 3 introduced a generic-bin condition in which participants estimated the chances of an observation being in the most common category, whatever it happens to be. We compared this with a specific-bin condition in which participants estimated the chances of an observation being in the specific category they identified as being most common. Whereas uncertainty is primarily aleatory for generic bins, specific-bin judgments include both aleatory uncertainty and epistemic uncertainty about which category is most common. We measured participants' confidence in their ability to identify the most common category in both conditions. Notably, average confidence was only slightly above the midpoint of the scale, so certainty about knowing the distribution is an unlikely explanation for the overprecision that we find.

Consistent with statistical prescriptions, specific-bin estimates increased with increasing confidence and generic-bin estimates much less so. These results tell us that people do recognize a difference between these two types of judgments, and they have intuitions about these judgments that align with Bayesian prescriptions in important ways. At the same time, judges with high confidence (low epistemic uncertainty) gave subjective probabilities that were more concentrated than estimates about the distribution. This excess concentration is a basic violation of statistical principles.

The results of Experiment 3 provide a possible explanation for why the manipulations of epistemic uncertainty in the first two experiments showed little effect. Participants who learned the median or the mode of the distribution were in a position analogous to the generic-bin condition of Experiment 3. Participants who did not receive that information were like those in the specific-bin condition. We did not elicit confidence judgments in Experiments 1 and 2. However, we can speculate that participants who were not told the mean or median varied in their confidence in their guesses. If high confidence leaked

into the probability assessments of the most confident individuals, as in the specific-bin condition of Experiment 3, their excessive concentration may have offset the lowered concentration of those who were less confident.

9. Experiment 4

The results of Experiment 3 show that participants were sensitive to epistemic uncertainty in both normatively appropriate and inappropriate ways, and we designed Experiment 4 to further probe these effects. In particular, we identified three different possible explanations, not mutually exclusive, for the sensitivity to epistemic uncertainty and the excess concentration found in Experiment 3. We express these possibilities as three new hypotheses.

(H4) *Confidence Leakage*: When forming a subjective probability distribution (SPD) about specific bins, people take both epistemic and aleatory uncertainty into account, but not in the prescribed way. Instead, confidence about the empirical distribution "leaks" into SPDs so that when epistemic uncertainty is high, judges' SPDs are less concentrated than they expect the empirical distribution to be, but when uncertainty is low, their SPDs are more concentrated than the expected empirical distribution.

(H5) *Rising Confidence Lifts All Bins*: When judges are more confident about the empirical distribution, they estimate higher probabilities for all categories, not just for those they deem most likely. This can push up the probability of the favored category, possibly beyond logical limits, without changing the concentration of the SPD as a whole. This violates the logic that all the probabilities in a mutually exclusive and exhaustive set of categories must sum to 1. However, violations of this principle have been observed in studies of "unpacking" (Koehler, Brenner and Tversky 1997; Rottenstreich and Tversky 1997). Those studies document subadditivity, in which the subjective probability of a class of events is lower than the sum of subjective probabilities assigned to the individual subcategories of which the class is

composed. A lifts-all-bins tendency would appear in raw probability judgments (Experiment 3) but would be statistically eliminated when SPDs are normalized to sum to 1. (Experiments 1 and 2).

(H6) *Single Bin Is Special*: There is something special about thinking about a single best category versus thinking about a whole distribution. The sensitivity to epistemic uncertainty and the excess concentration effects in Experiment 3 happen most when thinking about the single, top category. This explanation aligns with prior research on anchoring (Block and Harper 1991) and on the tendency to favor a single, focal hypothesis (Brenner and Koehler 1999).

As in Experiment 3, the current study includes estimates of specific-bin probabilities and of generic-bin probabilities. We expand upon the previous studies by eliciting judgments for multiple bins covering the full range of possible values (as in Experiments 1 and 2) in addition to judgments about the single, middle bin (as in Experiment 3). Epistemic uncertainty is measured via self-reported confidence. We also manipulated uncertainty in case judges who express high vs. low confidence differ in ways other than just their self-assessed knowledge about the city in question. Rather than providing or withholding information about the empirical distribution (as in Experiments 1 and 2), epistemic uncertainty was varied by asking questions about cities with which the participant is likely to be less or more familiar (out-of-state vs. in-state). As shown in Table 2, Hypotheses H4, H5, and H6 each imply a different pattern of results with respect to type of judgment (specific bin or generic bin), elicitation (full range or single bin), and level of confidence.

9.1 Method

9.1.1. Participants. Our pre-registered research plan called for a sample of 2,400 MTurk participants residing in the United States, roughly 300 participants in each of 8 conditions. Our analyses include data from 2,354 participants with an average age of 41.5 (*SD* = 12.5); 48% were female. We excluded 216 who failed the pretest, 55 who appeared to be duplications, and 82 whose completion times

Table 2. Predictions for three potential explanations for the results of Experiments 1 – 3.

	Middle-bin raw probability	Middle-bin normalized probability[1]	Mean Cumulative Confidence[1]	Sum of raw probabilities[1]
H4: Confidence Leakage	SP_M increases with confidence, stronger effect for SE than GE; $SP_{M, SE} > SP_{M, GE}$ at high confidence			No effect of confidence
H5: Rising Confidence Lifts All Bins	Same as for Confidence Leakage	SP_M similar in SE and GE little effect of confidence		Increases with confidence
H6: Single Bin is Special	**Single-bin elicitation:** Same as for Confidence Leakage **Full-range elicitation:** SP_M similar in SE and GE little effect of confidence	SP_M similar in SE and GE little effect of confidence		No effect of confidence

Note: SE = specific-bin probability estimates; GE = generic-bin probability estimates.
[1] These variables are available only with the full-range elicitation format.

were implausibly fast. In addition to a standard base payment of $1.25, participants received an average bonus of $0.79 for accuracy.

9.1.2. Design. The experiment followed a 2 x 2 x 2, between-subjects design. The three independent variables were judgment type (specific- vs. generic-bin), elicitation format (full range vs. single bin), and target city (in-state vs. out-of-state). To assure that in-state and out-of-state targets were comparable, each out-of-state participant received the city and question order given to one of the in-state participants.

Two of the dependent variables describe the entire distribution. These are the overall concentration measure SP_M and the sum of the raw probabilities across all bins. Two others are measures of the middle bin: the raw middle-bin probability and the normalized middle-bin probability. Normalized probabilities, concentrations, and sums of probabilities are calculated using the full range of estimates. Thus, raw middle-bin probability is the only measure available for single-bin elicitations. (Because the different explanations shown in Table 2 do not make distinct predictions regarding overprecision we do not discuss the O_M measure in detail here. See §S1.4 of the supplement for the analyses of O_M results.)

Each participant estimated temperatures for three different months (February, June, and October)

for their target city. We had no hypotheses concerning month-to-month differences and responses on each of the three months were highly correlated for all four dependent variables (r = .64 to .79 across all participants). Accordingly, the dependent variables in our analyses are the averages for each judge across the three months.

9.1.3 Materials and procedure. At the start of the survey, participants indicated the U.S. state in which they resided. Next was a five-question test of basic numeracy, followed by a brief training to help participants understand the elicitation procedures. Subsequently, there were three items in which we briefly described the National Weather Service data we used and asked for estimates about temperatures in a one city for three different months—February, June, and October—in a random order. For half of the participants the target city was the largest city in their home state; that city was subsequently assigned to a yoked participant in a different state. Participants indicated their estimates with sliding scales, as in the previous studies.

We modified the elicitation methods from Experiment 3 to make the specific bin and generic bin elicitations more similar. For generic-bin estimates, the median was described as "the median high temperature across all the days in June [or February or October] for the years 2017 to 2021, whatever it happens to be." The middle bin was labeled as "within 5° of the recorded June median." This was the only bin in single-bin elicitations. For full-range elicitations the other bins were "more than 15° cooler than the June median"; "between 6° and 15° cooler…"; "between 6° and 15° warmer…"; and "more than 15° warmer…"

For specific-bin estimates, prior to making their estimates for each month, participants were asked to provide their "best guess of the median high temperature in [city], across all days in [month]." That number was included in the label of the middle bin, such as "Within 5° of your guessed June median of 70 °F." For full-range elicitations the other bins were labeled as "more than 15° cooler than your guessed June median," etc.

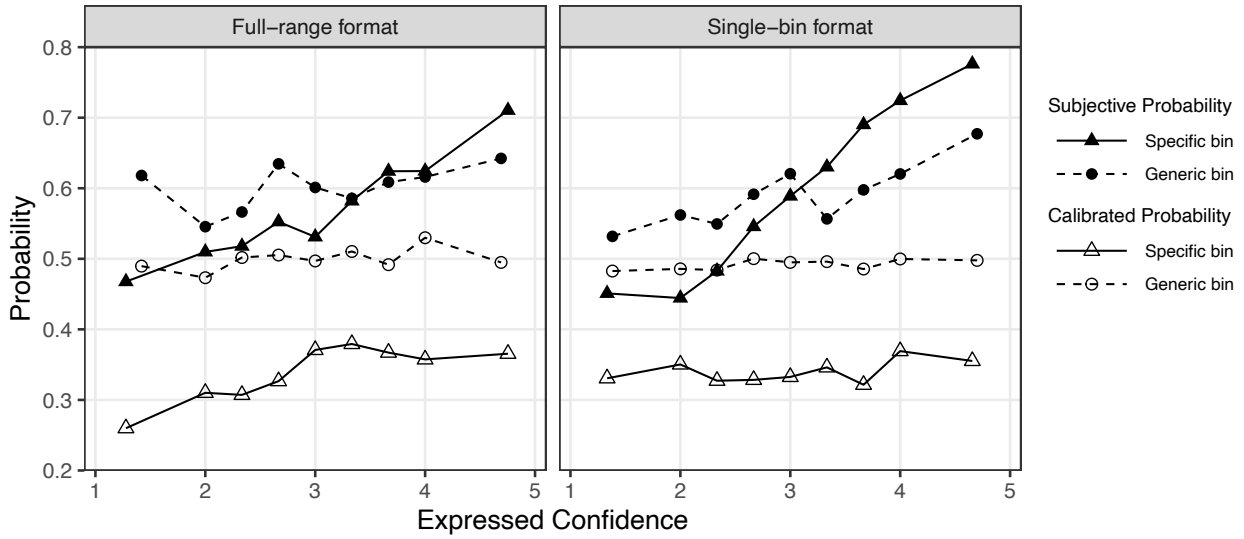
After completing their estimates for all three months in the given city, participants indicated their degree of confidence in their ability to estimate the median high temperature for each month. Those who had made specific-bin estimates were reminded that they had previously guessed the median high temperature in their target city for that month. Those who had made generic-bin estimates were asked to make that guess now. All participants were then asked, "How confident are you that your guess of [D] ^{o}F is within 5 degrees of the correct answer?", with their best-guess number inserted at [D]. They responded by choosing one point on a 5-point scale from "Not at all confident" to "Extremely confident." Finally, participants were asked optional questions about their gender and age.

9.2 Results

9.2.1 Raw middle-bin probabilities. This measure is analogous to the main dependent variable in Experiment 3, shown in Figure 5. We expected to replicate those results: an overall tendency to be overprecise, a main effect of confidence such that the probability assigned to the middle bin increases with increased confidence, a confidence x judgment interaction in which specific-bin estimates are more sensitive to confidence than generic-bin estimates are, and specific-bin estimates that are lower than generic-bin estimates when confidence is low and higher than generic-bin estimates when confidence is high.

We regressed the raw middle-bin probability judgments on Confidence, Judgment Type (coded as 1 = specific bin, -1 = generic bin), Format (1 = middle bin only, -1 = full range), and their interactions. As shown in Figure 6 and Table 3, all the predicted effects from Experiment 3 replicated. Spotlight analyses at ±1.5 standard deviations from the mean confirm that specific-bin estimates are lower than generic-bin estimates when confidence is low in both the single-bin format (t = 4.368, p < 0.001) and in the full-range format (t = 5.407, p < 0.001). Most crucially for the "Confidence-leakage" hypothesis, specific-bin estimates are higher than generic-bin estimates when confidence is high for both the single-bin format (t = 6.242, p < 0.001) and full-range format (t = 2.291, p = 0.022).

Figure 6. Raw estimates of middle-bin probabilities in Experiment 4.



Notes. *Expressed Confidence* is the confidence in identifying the median temperature of the target city, averaged across three months of the year. Probability corresponds to either subjective probabilities assigned to the middle bin or calibrated probabilities. Because of low *n* at the extremes of the confidence scale, these graphs combine the responses of judges with average confidence that is low (1, 1.33, and 1.67) and high (4.33, 4.67, and 5).

The results for the raw middle bin probabilities are consistent with both the "Confidence-leakage" and "Rising-confidence-lifts-all-bins" explanations, but not with the "Single-bin-is-special" explanation. As Figure 6 shows, the hypothesized pattern appears for both formats, whereas "Single-bin-is-special" implies that we would observe the pattern only with the single-bin format. The three-way interaction implied by the "Single-bin-is-special" explanation was not significant. Furthermore, the fact that the spotlights were significant separately for both formats provides additional evidence against this explanation.

Finally, overprecision in Figure 6 can be seen as the difference between the subjective and calibrated probability curves for a given condition. Overprecision was greater in the specific-bin condition, primarily because calibrated probabilities are less accurate when judging specific events, but also because of the confidence leakage effect. For a more detailed analysis of overprecision for all of our measures, see §S1.4 of the supplement.

Table 3. Regression analysis of dependent variables in Experiment 4.

	Raw Middle Bin		**Normalized Middle Bin**		**Concentration (SP_M)**	
Predictors	*Coef.*	*SE*	*Coef.*	*SE*	*Coef.*	*SE*
(Intercept)	0.597***	0.0038	0.349***	0.0037	0.683***	0.0023
Format	0.008*	0.0038				
Judgment	-0.003	0.0038	0.000	0.0036	-0.002	0.0023
Confidence	0.058***	0.0043	0.020***	0.0040	0.009***	0.0025
Format x Judgment	0.012**	0.0038				
Confidence x Format	0.016***	0.0043				
Confidence x Judgment	0.034***	0.0043	0.013***	0.0040	0.006*	0.0025
Confidence x Judgment x Format	0.007	0.0043				

Note: ** p<0.05 ** p<0.01 *** p<0.001.* Format = elicitation format, single-bin or full-range; Judgment = judgment type, specific-bin estimates or generic-bin estimates; Confidence = expressed confidence in estimates of the median temperature for the target city, averaged across judgments for February, June, and October. *N* = 2,354 for raw middle bin (R^2 = .098), *N* = 1,232 for normalized middle bin (R^2 = .026) and SP_M (R^2 = .014).

9.2.2. Normalized middle-bin probabilities and SP_M. According to the "Confidence-leakage" explanation, judges who are confident in their knowledge of the distribution also become confident about where in the distribution an observation will fall. This would affect both the probability estimates for their most likely category and the overall concentration of their SPD. In that case, we should see the same pattern of results for both the normalized middle bin probability and concentration as we did for the raw probability estimates. In contrast, "Rising-confidence-lifts-all-bins" and "Single-bin-is-special" both imply otherwise. Recall that because normalized probabilities and SP_M require a full distribution to calculate, these analyses can only use data from the full-distribution elicitation format.

Results for normalized middle-bin probabilities, shown in Table 3, indicate effects similar to those observed for raw middle-bin probabilities: a main effect of confidence and a confidence x judgment interaction showing that specific-bin probability estimates are more sensitive to confidence than are generic-bin estimates. In this regard, the results are consistent with Bayesian thinking. However, as before, a spotlight regression analysis shows that judges diverge from Bayes when they are highly confident; the predicted normalized middle-bin probabilities at +1.5 *SD* on the confidence scale are 0.393

vs. 0.358 for specific and generic bins, respectively, t = 2.69, p = .007. The results for SP_M, also shown in Table 3, are similar. The spotlight analysis for SP_M provides additional evidence that when confidence is high, concentration of SPDs is higher for specific bins than for generic, t = 1.66, p = .098. Confidence leakage may not be the only source of overprecision: Even those participants with low confidence were overprecise ($SP_M > CP_M$) when judging specific bins. At 1.5 SD below the mean confidence, overprecision was estimated to be 0.06 (t = 5.66, p < .001). Overall, however, the results for SP_M support the hypothesis of confidence leakage. For a more detailed analysis of overprecision, see §S1.4 of the supplement.

9.2.3. Sum of all probabilities. A direct implication of the "Rising-confidence-lifts-all-bins" explanation is that the sum of all probabilities assigned to the bins of the distribution will increase with increased confidence. In accord with others' findings regarding subadditivity, the sum of probabilities far exceeded 1 (t = 41.9, p < .001, d = 1.19), averaging 1.83. However, the total did not vary with confidence (t = 0.95, p = .344), nor was there an interaction between judgment type and confidence (t = 1.08, p = .279).

9.2.4 In-state vs. out-of-state cities. The in-state vs. out-of-state manipulation was intended to vary the degree of epistemic uncertainty. The manipulation was moderately successful. Mean expressed confidence for in-state cities was 3.30, SD = 0.98; mean expressed confidence for out-of-state cities was 3.04, SD = 1.02. An ANOVA showed that this difference was significant, $F_{1, 2346}$ = 49.13, p < .001, η_p^2 = .021 and did not differ significantly by judgment type or elicitation format.[6] We further tested the dependent variables with ANOVA models and were encouraged that the results were qualitatively similar, although weaker, to those seen with self-reported confidence. This is to be expected given the modest difference

[6] The ANOVA also revealed a main effect of judgment type: Judges in the specific-event condition expressed more confidence in their ability to guess the city's median temperatures, M = 3.28, SD = 0.98, than those in the generic-bin condition, M = 3.07, SD = 1.03, $F_{1, 2346}$ = 33.9, p < .001, η_p^2 = .014). We had no hypotheses about this effect, but we speculate that it is due to the difference in the timing of the guesses (interspersed with probability and distribution judgments or in the later block of confidence judgments).

in average confidence between the in-state and out-of-state groups, and the heterogeneity of confidence levels within each group (see §S1.4 of the supplement for full results).

9.3 Discussion

The results of this study confirm that people's intuitions align with Bayesian norms in an important sense: The greater their epistemic certainty, the more concentrated their SPDs. However, there is also an important way that intuition can lead people astray. When people are confident in their knowledge of the distribution, they may have SPDs that are more concentrated than they expect the population distribution to be. The best explanation for this excess concentration seems to be what we refer to as confidence leakage. That is, people lack the intuition that, normatively, the effects of epistemic uncertainty are strictly unidirectional.

An additional finding of Experiment 4 is that if people are not explicitly instructed to make their subjective probabilities sum to 100%, they will add up to more (nearly 200% in this case). The experiment cannot determine whether people held exaggerated confidence in each category as they went along, were not paying attention to the running total, or failed to infer that their estimates were supposed to sum to 100%. This is not a sufficient explanation for excess concentration, because when we normalize totals to 100%, we still see the effect. However, it may contribute to overconfidence in other ways. For instance, there are hints that excess concentration may be greater when thinking about a most likely outcome (when total sum of probabilities cannot be measured) than when thinking about the whole range of possible outcomes.

Finally, the results also reveal a discrepancy—middle bins exhibited substantially greater overprecision in Experiment 4 (0.29 and 0.10 for specific and generic bins, respectively) than observed in Experiment 3 (0.13 and 0.01). These results might be accounted for by differences in elicitation procedures; e.g., building a distribution around a guessed median, the use of sliders instead of numbers, and narrower category ranges in the stimuli. We leave investigation of these factors for future research.

10. General Discussion

Decision makers are regularly confronted with problems involving variables about which they are uncertain. To make well-reasoned judgments, they must draw on their imperfect knowledge and beliefs about these variables to quantify the likelihood of different outcomes. Often, this requires taking account of both epistemic and aleatory uncertainty.

Most prior work on subjective confidence has focused only on epistemic uncertainty. These prior studies use questions that have a unique correct value (e.g., the year in which Mozart was born). Uncertainty arises only from the judge's own lack of knowledge. However, people also face aleatory uncertainty, that is, variability due to stochastic processes (e.g., how long it will take to get to work today). A few studies have included questions that involve both uncertainties. For example, Soll and Klayman (2004) included a question about the winning percentage of a basketball team described by several characteristics, and Moore et al. (2015) asked for predictions about outcomes of several processes that people thought of as random. However, these studies were not designed to distinguish the effects of people's thinking about epistemic and aleatory uncertainty and their combination. Our work provides insights into these processes and the roles they play in overconfidence.

10.1 Empirical contributions

Judges often recognize that there is a distribution of values within any class of items or events, be it commute times or temperatures or prices. However, their knowledge about the distribution—its location, shape, and width—is almost never perfect. And, of course, judges do not know exactly what their errors may be. Because of this epistemic uncertainty, well-calibrated judges should have subjective probability distributions that are wider than they think the empirical distribution is. Their SPDs must account for both the distribution of possible values and their uncertainty about what that distribution is.

Moore et al. (2015) report that SPDs are less concentrated than the underlying empirical distributions, yet still overprecise. That is, SPDs were too concentrated to be well-calibrated given how

inaccurately their probabilities were placed. Looking at a variety of more familiar domains, our results suggest that people may be even further off the mark. We found that the difference between empirical and subjective distributions varied considerably from one domain to another, and on average, SPDs were slightly *more* concentrated than the corresponding empirical distributions. In every domain we studied, judges were overprecise, often by a wide margin.

We examined several different explanations for why this is the case. Across studies, we varied the level of epistemic uncertainty, sometimes quite transparently, and we provided hints and cues to bring epistemic uncertainty to front of mind. None of those manipulations had any appreciable net effect. The key word there is "net." People's response to epistemic uncertainty is not merely insufficient: They have a qualitatively incorrect intuition about how to incorporate it. They do correctly spread out their distributions when epistemic uncertainty is high, as prescribed by the laws of probability. However, when epistemic uncertainty is low, their SPDs are more concentrated than they think the empirical distribution is. Bayesian reasoning tells us this should not happen. The fact that it does, systematically, suggests that confidence about shape and location of the empirical distribution is confounded with confidence about which event in the distribution will be observed. We call this confidence leakage.

We do not mean to imply that confidence leakage is the only mechanism behind overprecision. Prior studies of SPDs with only epistemic uncertainty find a tendency to overestimate one's knowledge. Moreover, when uncertainty is framed or perceived in more epistemic terms, probability judgments tend to be more extreme, leading to greater overprecision (Tannenbaum et al. 2017). And we find overprecision even in judges who express high epistemic uncertainty. Such findings suggest insufficient response to epistemic uncertainty is also part of the picture. In the introduction, we proposed three possible causes of insufficient adjustment: believing that events in the world are more concentrated than they really are, believing that you know the distribution better than you really do, and spreading your SPDs incorrectly given your level of knowledge. Our studies cannot settle this, but Studies 3 and 4 do offer

relevant evidence. Judges' subjective distributions were not systematically more or less concentrated than the corresponding empirical ones. Judges who expressed very low confidence were unlikely to be underestimating their uncertainty, but were overprecise nonetheless. Thus, subject to further research, the third mechanism seems most likely. People are overprecise because they respond to epistemic uncertainty both insufficiently and inappropriately.

Our results recall the work of Griffin et al. (1990), who found that people neglect epistemic uncertainty regarding their underlying assumptions. For example, confidence intervals for spending on an evening out were the same whether or not participants knew relevant details (e.g., which restaurant, what food was ordered, etc.). Although drawing participants' attention to these uncertainties succeeded in expanding confidence intervals, by default they acted as though their construals of the situation were accurate. Our findings mirror and expand upon these results. For instance, we find that whereas people do expand their distributions in the face of epistemic uncertainty about the distribution, they also concentrate them when they are certain. This confidence-leakage effect represents yet one more factor that contributes to overprecision.

10.2 Methodological contributions

We introduce a new method for characterizing the concentration of a probability distribution that allows a researcher to evaluate both (a) the concentration of a subjective distribution relative to the empirical distribution and (b) whether the subjective distribution is overprecise—too concentrated to achieve good calibration. These assessments rely on comparisons between mean cumulative concentration (MCC) measures, which describe the extent to which the probabilities from a particular distribution coalesce around a specific set of outcomes rather than being spread evenly across all outcomes. By comparing the judged, empirical, and calibrated MCCs of a particular distribution, researchers can study the sources of overprecision in ways that are not possible using existing metrics such as absolute deviations and interval hit rates. Furthermore, all MCCs are measured in units of

probability, thereby allowing for standardized comparisons across different distributions with different units, which facilitates analysis of judgments from a variety of domains.

10.3 Limitations and future directions

Our online samples provided greater diversity and larger sample sizes than would be possible in a laboratory. However, as with any diverse sample, there is room to be concerned about participants' numerical sophistication and their motivation to be accurate. We attempted to address concerns about numerical sophistication by screening participants with tests of numeracy. We cannot be certain that the screening ruled out all misunderstandings of necessary numerical concepts, although we think it unlikely that our sample was appreciably less numerate than the general population. We attempted to motivate accuracy by providing monetary incentives that rewarded accurate responses. Although the size of these incentives was in line with recent norms for online participants, they might still have been insufficient to ensure strong accuracy motivation. It would be interesting to learn more about how incentives and mathematical training might affect subjective confidence judgments.

Like much of the prior literature, our experiments utilize assessments of probability distributions and numerical probabilities. It has long been established that ordinary people misunderstand numerical probabilities (e.g., Fischhoff 1991), and it is unlikely that many people naturally think of uncertainty in terms of probability distributions. Thus, the elicitation methods we, and many previous investigators, use are unfamiliar to participants. Given how many decisions in life, from investment to clothing choices, depend on understanding probability and hedging risks, important questions remain about whether behavioral measures of certainty, such as choices under risk, might reveal more accurate intuitions (Mamassian, 2008; Mannes and Moore, 2013), and whether domain experts familiar with assessing risk are susceptible to the same errors when combining aleatory and epistemic uncertainty.

We have focused in this paper on the difficulty people have in combining aleatory and epistemic uncertainty. People have some flexibility in framing uncertainty as one or the other, and the effects of

that framing are worthy of future research. For example, people can learn probabilities by observation rather than by a summary description of outcomes. Studies find that overprecision is lower with direct observation of data (Budescu and Du 2007, Camilleri and Newell 2019). Learning from observation may favor an aleatory representation of the problem, which, based on our findings, should reduce overprecision. Future research might also investigate the implications for the accuracy-informativeness tradeoff (Yaniv and Foster 1995), by which people avoid very wide confidence intervals because they are unhelpful to the listener (e.g., "90% confident that travel time is between 3 hours and 5 days"). Assessments of uncertainty may be more elastic when represented as epistemic rather than aleatory and therefore more likely to favor informativeness over accuracy.

10.4 Conclusions

Our work touches on fundamental questions about how people know what they know. We find that people's estimates about a distribution of possible outcomes are systematically different from what normative rules prescribe. Their probability judgments are overprecise, meaning that people underestimate the magnitude of their errors (Soll and Klayman 2004). Subjective probability distributions do not properly reflect the degree of (in)accuracy in the judge's knowledge. We can easily imagine our participants objecting to our characterization of them. How could we expect them to know what they do not know? And yet, an appropriate level of confidence requires the application of exactly that kind of metacognition. How to do so is not at all obvious. Statistical reality demands that uncertainty in the placement of a distribution widen the distribution of possible outcomes, but this reality is not intuitively obvious to most people, at least not to our participants. As long as there are things people do not know, it will be difficult for them to properly take into account their lack of unknown information when calibrating their confidence judgments (Moore 2022). Conditions on the ground rarely permit perfect calibration, and people are unlikely to get the unbiased, timely, and plentiful feedback needed to come very close (Hogarth et al. 2015). Extant evidence shows that overprecision predominates in judgments

about any range of outcomes. Our analyses reveal that a likely contributor to that error is a misunderstanding about how epistemic and aleatory uncertainty combine. Overconfidence in subjective probability distributions poses a key challenge for decision making in management, healthcare, personal finance, and myriad other domains. We hope that our work contributes to the development of effective methods to bring subjective confidence more in line with reality.

11. References

Alpert M, Raiffa H (1982) A progress report on the training of probability assessors. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. (Cambridge University Press, Cambridge, MA), 294-305.

Baillon, A, Placido, L (2019) Testing constant absolute and relative ambiguity aversion. *J. Econ. Theory*. 181:309-332.

Ben-David I, Graham, JR, Harvey, CR (2013) Managerial miscalibration. *Q J Econ. 128*(4):1547-1584.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. Organizational behavior and human decision processes, 49(2), 188-207.

Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: Local-weight models for decomposition of evidential support. Cognitive Psychology, 38(1), 16-47.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* 78(1):1-3

Boston Athletic Association. Champions of the Boston Marathon: Men's Open Division. Accessed January 26, 2023, https://www.baa.org/races/boston-marathon/results/champions.

Budescu DV, Du N (2007) The coherence and consistency of investors' probability judgments. *Manag. Sci.* 53(11):1731–1745.

Camilleri AR, Newell BR (2019) Better calibration when predicting from experience (rather than description). *Organ. Behav. Hum. Decis. Process.* 150:62–82.

Dawid, AP (1982) The well-calibrated Bayesian. *J. Am. Stat. Assoc*. 77(379):605-610.

Ellsberg, D (1961). Risk, ambiguity, and the Savage axioms. *Q J Econ*, 75(4):643-669.

Epley N, Dunning D (2006) The mixed blessings of self-knowledge in behavioral prediction: Enhanced discrimination but exacerbated bias. *Pers. Soc. Psychol. Bull. 32*(5):641–655.

Fischhoff B (1991) Value elicitation: Is there anything in there? *Am. Psychol.* 46(8):835–847.

Fox CR, Ülkümen G (2011) Distinguishing two dimensions of uncertainty. Brun W, Keren G, Kirkebøen G, Montgomery H, eds. *Perspect. Think. Judg. Decis. Mak.* (Universitetsforlaget, Oslo), 21–35.

Glaser M, Weber M (2007) Overconfidence and trading volume. *Geneva Risk Insur. Rev.* 32:1–36.

Goldstein DG & Rothschild D (2014) Lay understanding of probability distributions. *Judgm. Decis. Mak.*, 9(1):1-14.

Griffin, D. W., Dunning, D., Ross, L. (1990) The role of construal processes in overconfident predictions about the self and others. *J. Pers. Soc. Psychol.,* 59(6): 1128-1139.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgm. Decis. Mak.* 5(7):467–476.

Hogarth, RM, Lejarraga, T, Soyer, E (2015) The Two Settings of Kind and Wicked Learning Environments. *Curr. Dir. Psychol. Sci.*, 24(5), 379–385.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. *Management Sci.* 59(9):1970-1987.

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.

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: A cognitive perspective on risk and risk taking. *Management Sci.* 39(1):17–31.

Kahneman, D., Tversky, A. (1982) Variants of uncertainty. *Cognition*, 11(2):143-157.

Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Hum. Decis. Process.* 79(3):216–247.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997). The enhancement effect in probability judgment. Journal of Behavioral Decision Making, 10(4), 293 -313.

Lichtenstein SB, Fischhoff B, Phillips, LD (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. (Cambridge University Press, Cambridge, MA), 306-334.

Mamassian P (2008) Overconfidence in an objective anticipatory motor task. *Psychol. Sci.* 19(6):601–606.

Mannes AE, Moore DA (2013) A behavioral demonstration of overconfidence in judgment. *Psychol. Sci.* 24(7):1190–1197.

Moore, DA (2022) Overprecision is a property of thinking systems. *Psychol. Rev*. https://doi.org/10.1037/rev0000370

Moore DA, Carter A, Yang HHJ (2015) Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organ. Behav. Hum. Decis. Process.* 131:110–120.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychol. Rev.* 115(2):502–517.

Nisbett RE, Krantz DH, Jepson C, Kunda Z (1983) The use of statistical heuristics in everyday inductive reasoning. *Psychol. Rev.* 90(4):339–363.

Nisbett RE, Kunda Z (1985) Perception of social distributions. *J. Pers. Soc. Psychol.* 48(2):297.

Paté-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of treatment. Reliability Engineering & System Safety, 54(2-3), 95-111.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. Psychological Review, 104(2), 406 -415.

Russo JE, Schoemaker PJH (1992) Managing overconfidence. *MIT Sloan Manag. Rev*. 33(2):7-17.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Exp. Psychol. Learn. Mem. Cogn.* 30(2):299–314.

Tannenbaum D, Fox CR, Ülkümen G (2017) Judgment extremity and accuracy under epistemic versus aleatory uncertainty. *Management Sci.* 63(2):497–518.

Teigen KH, Jørgensen M (2005) When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Appl. Cogn. Psychol.* 19(4):455–475.

Ülkümen G, Fox CR, Malle BF (2016) Two dimensions of subjective uncertainty: Clues from natural language. *J. Exp. Psychol. Gen.* 145(10):1280–1297.

Yaniv I, Foster DP (1995) Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *J. Exp. Psychol. Gen.* 124(4):424–32.

Supplementary material for:

Overconfidence in probability distributions: People know they don't know but they don't know what to do about it

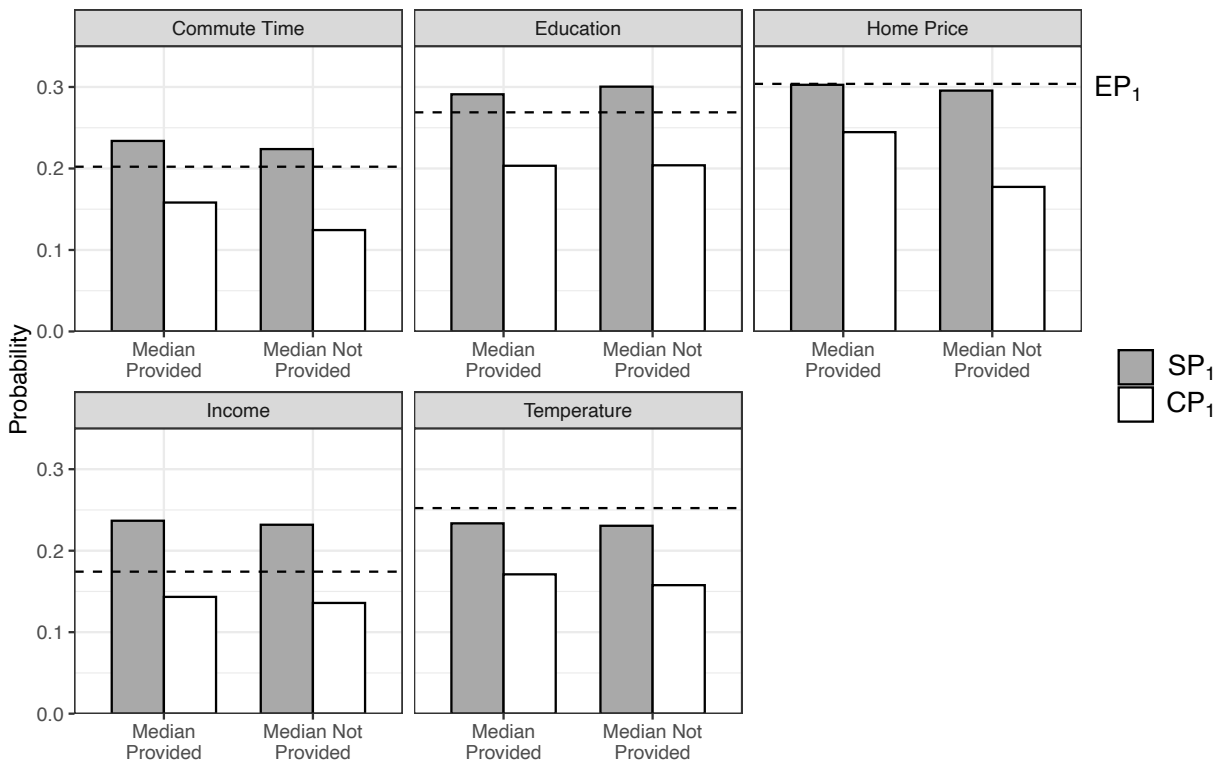
Contents

S1. Supplementary Analyses of Experimental Data

S1.1. Experiment 1.

In the main text, we analyzed the overall concentration of the distributions as measured by SP_M, EP_M, and CP_M. Here, we present results for the probability in the most likely category as measured by SP_1, EP_1, and CP_1. These results are displayed in Figure S1. Participants assigned probability SP_1 to the category they thought was most likely. The dashed lines denote EP_1, the probability of the most likely category in the empirical distribution. CP_1 is the empirical probability of the category the participant thought was most likely.

Figure S1. Average most-likely-category probabilities by variable domain in Experiment 1.



We analyze SP_1 and OP_1 using mixed-model ANOVAs with Information (median provided vs. not provided) as a between-subjects variable and Domain as a within-subjects variable. The grand means for the three concentration measures were $SP_1 = 0.258$, $EP_1 = 0.240$, and $CP_1 = 0.171$. Contrary to H1, and as reflected in the mean difference score, participants' SPDs were significantly more concentrated than the empirical distribution ($SP_1 > EP_1$), $F_{1, 592} = 53.3$, $p < .001$, $\eta_p^2 = .083$. Furthermore, contrary to H2, participants' SPDs as measured by SP_1 were similarly concentrated regardless of whether they saw the median, $F_{1, 592} = 0.414$, $p = .520$.

In support of H3, participants were overconfident on all five domains, $SP_1 > CP_1$, reflecting the fact that participants assigned more probability to their favored category than the empirical distribution did to that same category. This was true even in the temperature domain because, although participants' judgments were less concentrated than the empirical distribution, they often favored a temperature category that was not, in fact, the empirically most likely one. Over all domains, participants were significantly overconfident, $M_{O_1} = 0.086$, $F_{1, 592} = 1114.6$, $p < .001$, $\eta_p^2 = .653$. However, they were significantly less overconfident when provided with the median compared to when they did not have it, $M_{O_1} = 0.076$ vs. 0.097, $F_{1, 592} = 16.7$, $p < .001$, $\eta_p^2 = .027$. As shown in Figure S1, this was a consequence of being more accurate with the median (CP_1 was 0.184 and 0.160 with and without the median, respectively) without being more concentrated (SP_1 was 0.260 and 0.256 with and without the median, respectively).

S1.2. Experiment 2.

In the main text, we analyzed the overall concentration of the distributions as measured by SP_M, EP_M, and CP_M. As evident in Figure S2, domains varied in whether participants were more or less concentrated than the empirical distribution.

Figure S2. Average MCC measures by condition and variable domain in Experiment 2.

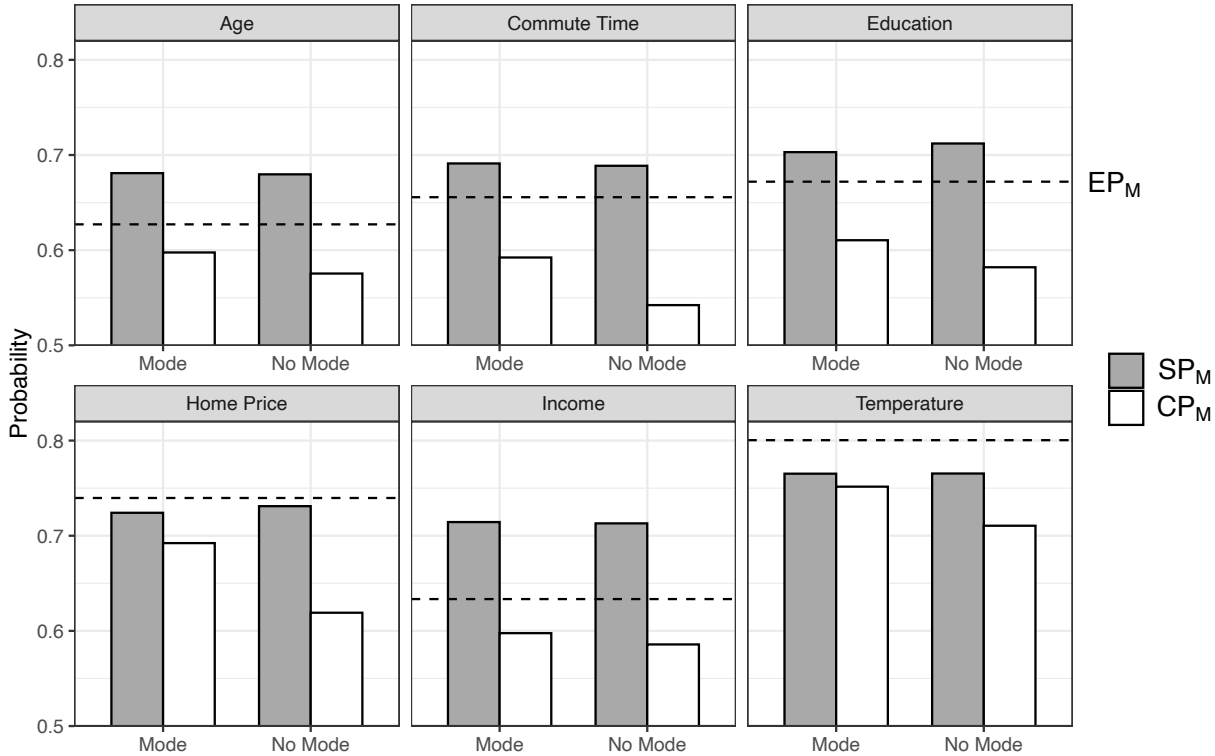
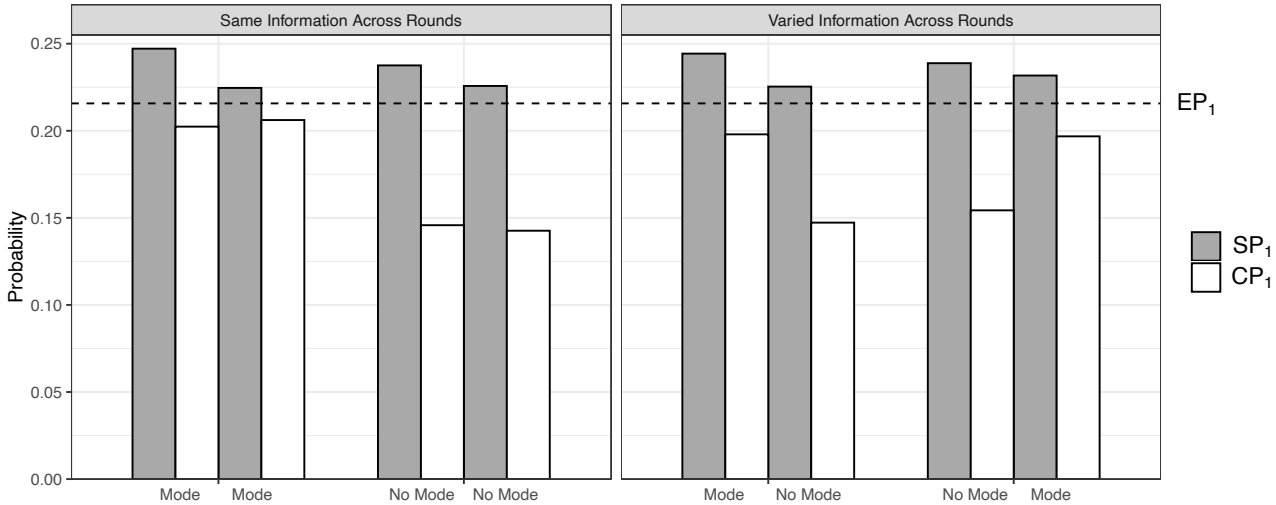


Figure S3. Concentration and calibration of most-likely-category probabilities in Experiment 2.



We also analyze concentration as measured by the most-likely-category probabilities SP_1, EP_1, and CP_1, displayed in Figure S3. As in our analysis of the MCC measures, we average across the three domains within each round, allowing us to analyze the data as a 4 (Information condition) X 2 (Round) mixed-model ANOVA. Because the design was perfectly balanced, the average EP_1 was 0.216 in all conditions and rounds. Overall, participants' distributions were more concentrated than the empirical ones: The mean difference between SP_1 and EP_1 was 0.019, $F_{1, 668} = 90.9$, $p < .001$, $\eta_p^2 = .120$.

An analysis of O_1 also reveals overconfidence, $M_{O_1} = 0.060$, $F_{1, 668} = 892.2$, $p < .001$, $\eta_p^2 = .572$. Participants were less overconfident in the second round, $M_{O_1, R1} = 0.067$ vs. $M_{O_1, R2} = 0.054$, $F_{1, 668} = 24.1$, $p < .001$, $\eta_p^2 = .035$. This reduction in overprecision happened because participants were less concentrated, but not less accurate, in the second round (CP_1 was 0.175 and 0.173 in rounds 1 and 2, respectively). As was the case for the MCC measures, there was less overprecision when the mode was provided. This captured by a Round x Information interaction: The effect of round depended on the presence or absence of the mode in that round, $F_{3, 668} = 40.8$, $p < .001$, $\eta_p^2 = .155$.

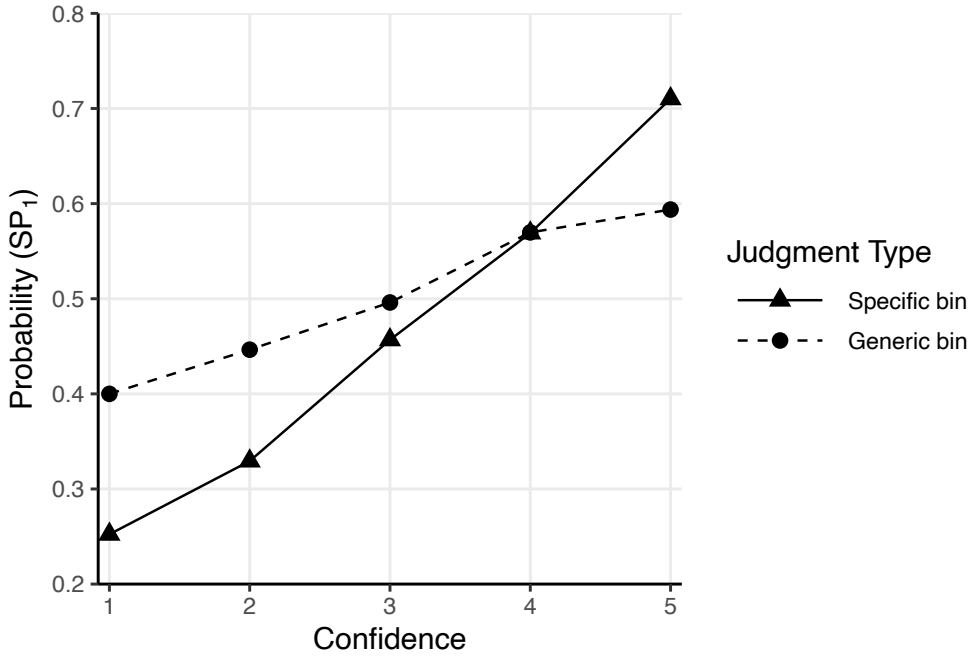
S1.3. Replication of Experiment 3.

Because the interesting results shown in Figure 5 of the main text were not among Experiment 3's preregistered hypothesis tests, we elected to test their reliability by replicating the key conditions from Experiment 3. We preregistered a plan (https://osf.io/mp5jb/) focusing on that

analysis. The plan required 600 MTurk participants, each randomly assigned to one of four experimental conditions: (1) Generic bin, with confidence prompt first; (2) Specific bin, with confidence prompt first; (3) Generic bin, with confidence prompt later; and (4) Specific bin, with confidence prompt later.

We planned to collect data in batches, examining them only to drop cases according to pre-specified exclusion criteria, until we achieved the planned sample size. That plan led us to collect data from 1186 individuals. We dropped 508 of these for scoring below 4 out of 5 on our numeracy quiz. We dropped an additional 32 who took less than 5 minutes to complete the study. We dropped 3 more responses that originated from duplicate IP addresses, raising fears that the same person could have participated more than once. That left us with a final sample size of 645.

Figure S4. Relationship between Confidence in identifying the most common category and the estimated Probability of a randomly chosen item falling into the category believed most likely (Specific bin) and or estimated Probability of a randomly chosen item falling in the most likely category, whichever category that is (Generic bin), in the replication of Experiment 3.

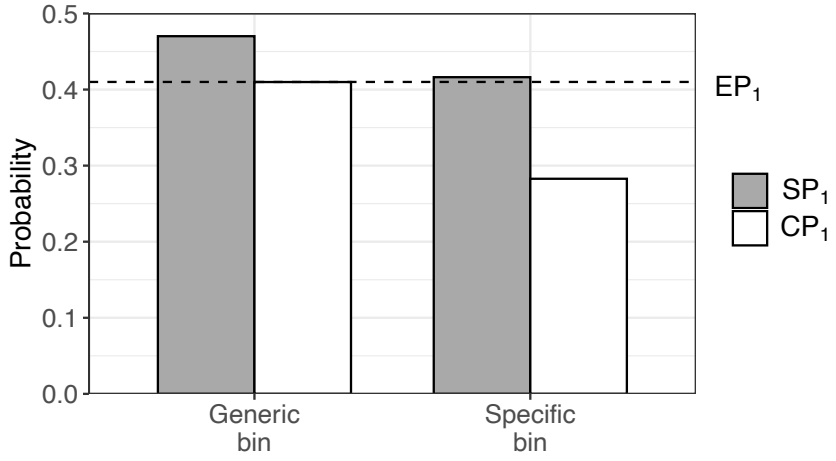


We expected to replicate Study 3's significant Judgment Type x Confidence interaction, such that percentage estimates increase with greater confidence, and more so in the Specific bin condition than in the Generic bin condition. We planned to test this with a regression analysis on estimates with Judgment Type (Generic vs Specific), Confidence, and Domain as independent variables, along with the Judgment Type x Confidence interaction, with robust standard errors

clustered by participant. As predicted, this analysis indicates that Estimates of probability in the Generic bin condition increased with greater confidence about which was the leading category, b = 0.050, S.E. = .008, t(644) = 6.38, p < .001. Estimates of probability in the Specific bin condition increased with confidence more so, b = 0.110, with a significant Type x Confidence interaction, S.E. = .011, t(644) = 5.33, p < .001. This pattern can be seen clearly in Figure S4.

These results replicate the effect identified in Experiment 3, wherein participants acknowledge that they are unsure about the values they are estimating, but do not know how to appropriately adjust their confidence judgments to account for that uncertainty. In particular, the probability assigned to the most common category should always have been lower for the Specific bin judgments than for the Generic bin judgments because the Specific bin judgments involve an additional source of uncertainty, over and above all the uncertainties present in the Generic bin judgments. When participants were maximally confident (a 5 on the 5-point Confidence scale) about which category was, in fact, the empirically most likely category, then they could be normatively justified in assigning as much probability to it in the Specific bin and Generic bin conditions. Instead, we again see that, confident participants in the Specific bin condition nevertheless assign higher probabilities than those in the Generic bin condition.

Figure S5. Concentration and calibration of the Generic bin and Specific bin judgments in the replication of Experiment 3.



Next, we study overconfidence as measured by O_1 using a mixed-model ANOVA with Judgment Type (Generic bin vs. Specific bin) as a between-subjects variable and Domain as a within-subjects variable. The mean overconfidence was 0.060 for Generic bin judgments and 0.134 for Specific bin judgments. As predicted, participants were significantly overconfident in their

judgments about the category that was most likely, whatever it happened to be, $F_{1,643} = 235.2$, $p < .001$, $\eta_p^2 = .268$, and even *more* overconfident in their judgments about the category that they *believed* was most likely, $F_{1,643} = 33.7$, $p < .001$, $\eta_p^2 = .050$, as evident in Figure S5.

S1.4. Experiment 4.

We test the middle-bin probability (both the raw and renormalized versions where possible) and SP_M using analyses of variance with judgment type, city type, and format (only possible for the raw middle-bin probabilities) as independent variables. These results are shown in Tables S1, S2, S3, and S4. In almost all cases, the results were qualitatively similar to, albeit weaker than, those seen when using self-reported confidence in the main text. The weaker relationship makes sense given the modest difference in average confidence between the in-state and out-of-state groups and the heterogeneity of confidence levels within each group.

Table S1. ANOVA of the raw subjective middle-bin probability as a function of condition.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	2346	2.11	0.147	0.001
Judgment Type (Generic vs. Specific)	1	2346	0.87	0.351	0.000
Format (Single-Bin vs. Full-Range)	1	2346	7.11	0.008	0.003
City Type x Judgment Type	1	2346	5.21	0.023	0.002
City Type x Format	1	2346	7.68	0.006	0.003
Judgment Type x Format	1	2346	14.20	0.000	0.006
City Type x Judgment Type x Format	1	2346	0.72	0.398	0.000

Table S2. ANOVA of normalized subjective middle-bin probability as a function of condition.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1228	8.03	0.005	0.006
Judgment Type (Generic vs. Specific)	1	1228	0.09	0.762	0.000
City Type x Judgment Type	1	1228	2.21	0.138	0.002

Table S3. ANOVA of SP_M as a function of condition.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1228	7.46	0.006	0.006
Judgment Type (Generic vs. Specific)	1	1228	0.22	0.643	0.000
City Type x Judgment Type	1	1228	4.97	0.026	0.004

Table S4. ANOVA of the raw subjective middle-bin probability in Own city by condition.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
Judgment Type (Generic vs. Specific)	1	1173	5.54	0.019	0.005
Format (Single-Bin vs. Full-Range)	1	1173	15.84	0.000	0.013
Judgment Type x Format	1	1173	11.40	0.001	0.010

For participants who provided judgments using the Full-Range format, we can also analyze the concentration and calibration of their SPD once normalized to a proper distribution with probabilities that sum to 1. Here, we calculate SP_1, EP_1, and CP_1 and SP_M, EP_M, and CP_M using the full normalized SPD. As can be seen in Figures S6 and S7, participants' SPDs were on average overprecise for Specific Bin judgments but close to well-calibrated for Generic Bin judgments.

Figure S6. SP_1, CP_1 and EP_1 by Judgment Type and City Type in the Full-Range Format in Experiment 4 using the full normalized SPD.

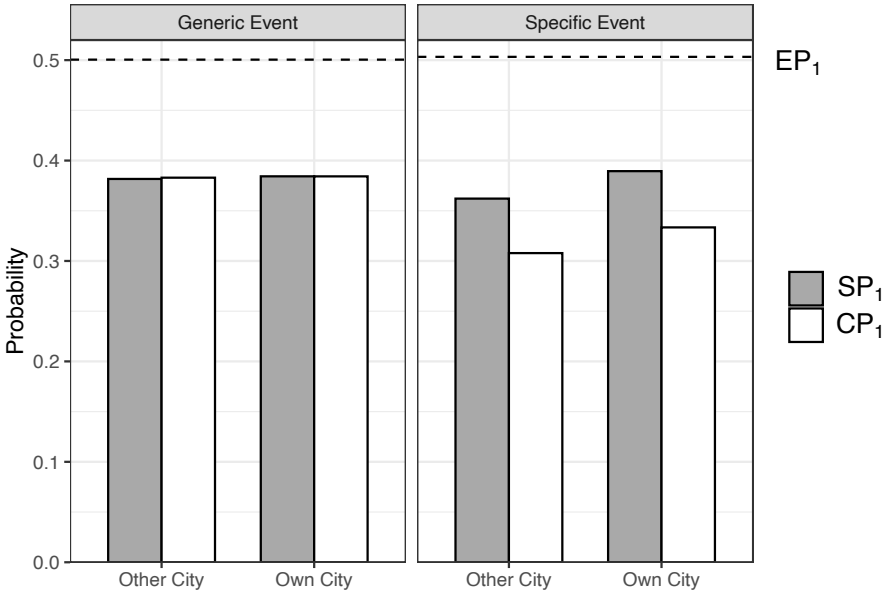


Figure S7. SP_M, EP_M, and CP_M by Judgment Type and City Type in the Full-Range Format in Experiment 4 using the full normalized SPD.

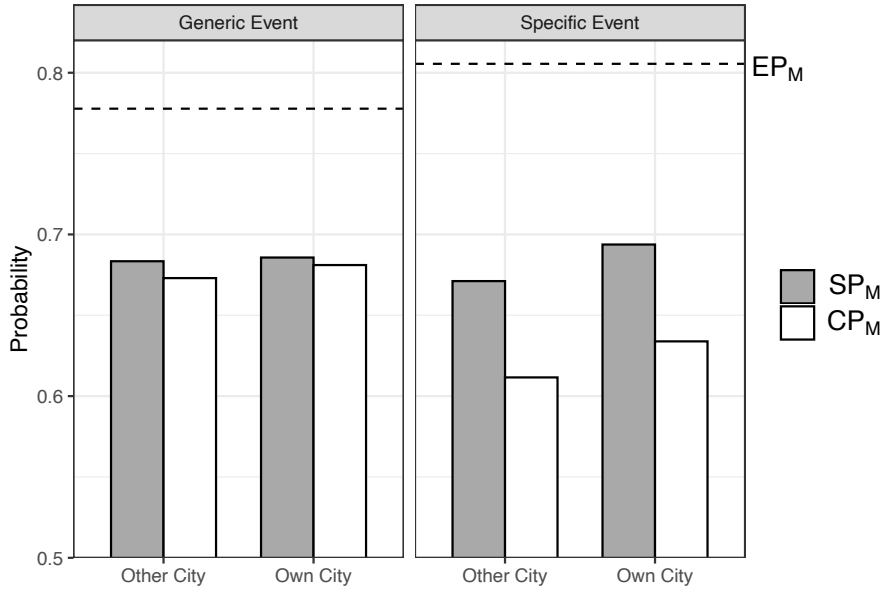


Table S5. ANOVA of overprecision of the raw subjective middle-bin probability in the Single-Bin format.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1118	4.13	0.042	0.004
Judgment Type (Generic vs. Specific)	1	1118	246.91	0.000	0.181
City Type x Judgment Type	1	1118	1.70	0.192	0.002

Table S6. ANOVA of overprecision of the raw subjective middle-bin probability in the Full-Range format.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1228	2.61	0.107	0.002
Judgment Type (Generic vs. Specific)	1	1228	97.47	0.000	0.074
City Type x Judgment Type	1	1228	0.01	0.916	0.000

Table S7. ANOVA of overprecision of the normalized subjective middle-bin probability in the Full-Range format.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1228	1.36	0.244	0.001
Judgment Type (Generic vs. Specific)	1	1228	272.71	0.000	0.182
City Type x Judgment Type	1	1228	0.01	0.932	0.000

Table S8. ANOVA of overprecision (as measured by O_M) of the full SPD in the Full-Range format.

Effect	**DFn**	**DFd**	***F***	***p***	η_p^2
City Type (Own vs. Other)	1	1228	0.15	0.700	0.000
Judgment Type (Generic vs. Specific)	1	1228	54.37	0.000	0.042
City Type x Judgment Type	1	1228	0.18	0.670	0.000

Next, we examine overprecision in Experiment 4. As shown in Figure 6 of the paper, the raw middle-bin probability judgments are consistently overconfident for judgments of both specific and generic bins. In all cases, the probabilities provided for the middle bins were on average higher than the actual fraction of outcomes that fell in those bins (the calibrated probabilities). In addition, participants who estimated specific bins exhibited both more overprecision overall and greater increases in overprecision with confidence, compared to those who estimated generic bins. This happened for two reasons. First, the calibrated percentages were lower in the specific bin conditions because participants' point estimates of the median were imperfect. Second, at high levels of confidence, participants who considered specific bins as opposed to generic ones estimated higher probabilities (the leakage effect).

Figure S8. Concentration measures and overprecision in Experiment 4.

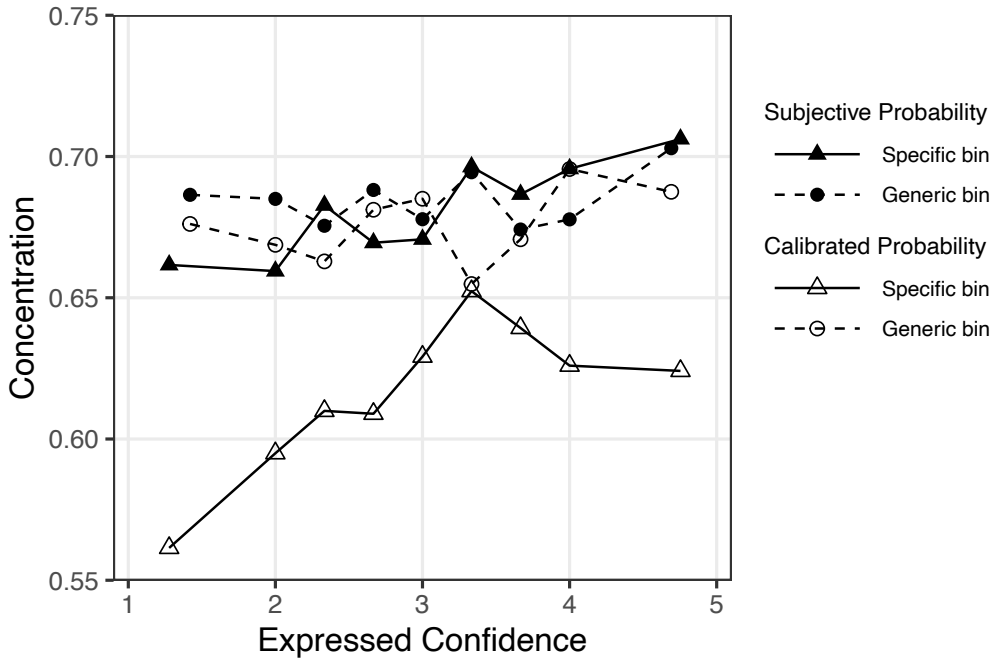


Figure S8 shows the results for overconfidence measured over the entire distribution (O_M), shown by the difference between subjective and calibrated probabilities. As with the raw single-bin estimates, participants were more overprecise when judging specific bins than generic bins (O_M = 0.060 vs. 0.008, t = 7.38, p < .001, d = .42, and more confident participants reported distributions that were more concentrated as measured by SP_M. In this case, however, overprecision when judging specific bins did not vary with confidence (r ≈ 0) because judges' accuracy, as measured by CP_M, increased more or less proportionately to increased concentration as confidence increased.

Figure S8 also indicates that even those participants with low confidence were overprecise ($SP_M > CP_M$) when judging specific bins. We confirmed this result by regressing overprecision ($O_M = SP_M - CP_M$) against expressed confidence centered at 1.5 SD below the mean, separately for the specific and generic bin conditions. The intercept in this analysis estimates overprecision at the centered "spotlight" value of 1.82. Overprecision was estimated to be 0.06 (t = 5.66, p < .001) for specific bins, which can be compared to 0.01 for generic bins (t = 1.21, p = .227). These results suggest that confidence leakage is not the only reason for overprecision.

S2. Additional Properties of the MCC Measures and Comparison with Other Measures.

In this section we explain how each of the three MCC measures (SP_M, EP_M, and CP_M) corresponds to a Gini coefficient and discuss how the measures compare to other, more-commonly used measures of the dispersion and calibration of a distribution. In doing so, we hope to provide evidence that the measures we use in our analyses in the main text of the paper largely reproduce the patterns that would be suggested by other candidate measures.

S2.1. Properties of the MCC measures and their connection to Gini coefficients.

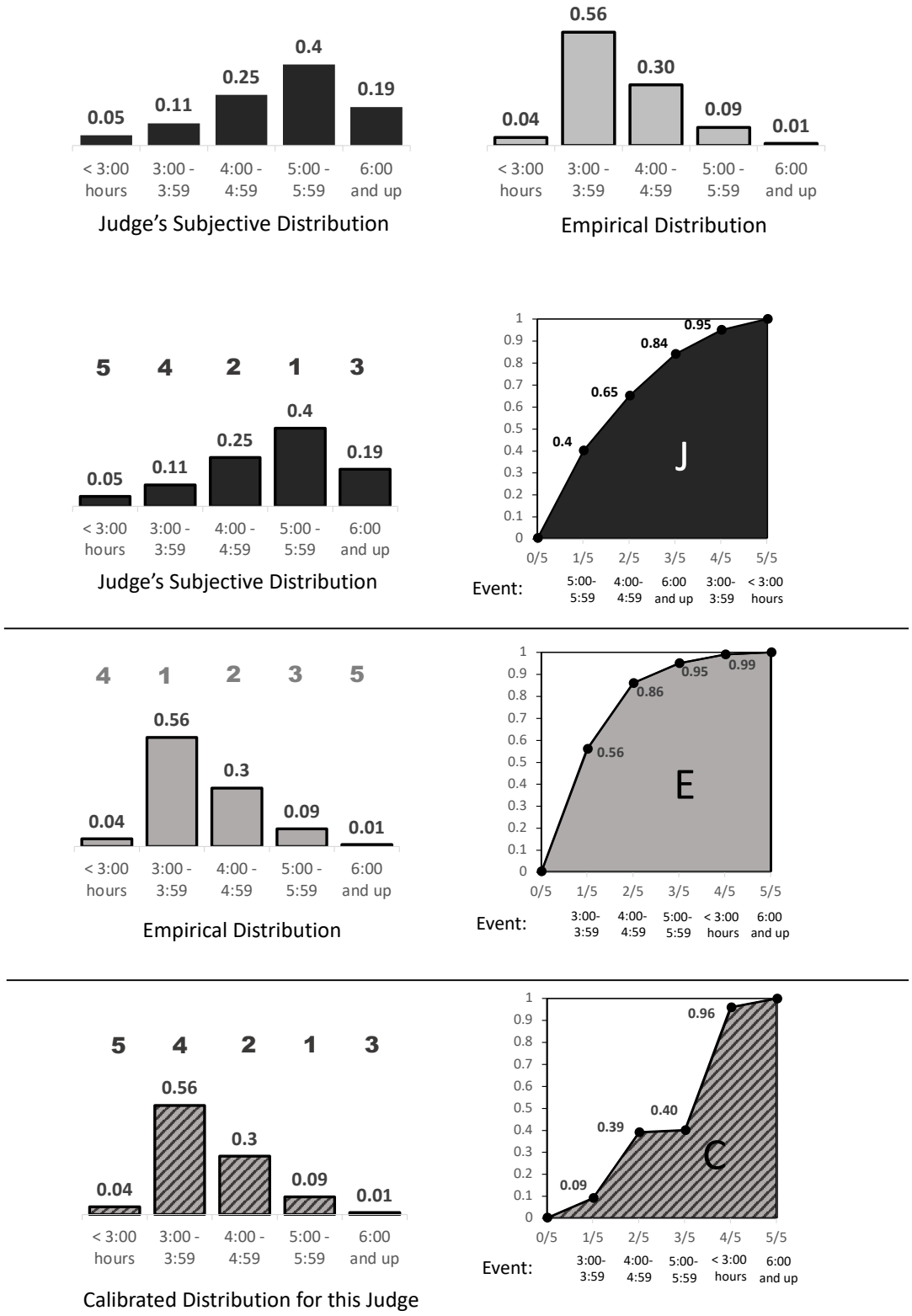
The MCC measures also correspond to a pair of well-established statistics used by population economists to measure the concentration of a resource across a population, namely the Lorenz curve and its numerical summary, the Gini coefficient (Gastwirth, 1972; Gini, 1912; Lorenz, 1905). The Gini coefficient is most commonly used to index the extent to which wealth is concentrated in a small subset of the population versus evenly distributed across individuals. Similarly, the MCC is an index of the extent to which probability is concentrated in the subset of categories deemed more likely versus spread out evenly across all outcome categories.

Figure S9 provides an example of Lorenz and Gini calculations for an individual who has provided probability judgments for the finishing time of an individual chosen at random from among all those running an upcoming marathon. The individual's responses for the $n = 5$ total categories appear in black in the upper left panel. This judge estimated finishing times between 5:00 and 5:59 hours to be most likely, assigning a subjective probability of 0.4 to this event, and estimating finishing times under 3:00 hours to be least likely, assigning this event a subjective probability of 0.04.

To generate a Lorenz curve, we subdivide the horizontal axis from 0 to 1 into equal fractional increments, with each increment adding another category cumulatively. Between those endpoints, the first increment represents the category assigned the most probability, the first two increments represent the *two* categories assigned the most probability, and so on. The notation 3/5, for example, indicates the calculation for the top three out of five total categories. The vertical axis tallies the corresponding cumulative probabilities assigned to each "top-k" subset of categories.

The Lorenz curve for the judged SPD, plotted in the top right panel, is constructed by successively cumulating the subjective probabilities of the ranges, starting with the one judged most likely (5:00 to 5:59) and ending with the one judged least likely (< 3:00). The steepness of

Figure S9. Calculation of Lorenz curves and Gini coefficients for a hypothetical judge estimating the distribution of marathon finishing times.



the judge's curve represents the degree to which the judge believes that some ranges are more likely than others. In the least concentrated possible distribution, each category is assigned equal probability and the curve follows the identity line from (0,0) to (1,1). The judge's Gini coefficient is given by $G_{judge} = \frac{n}{n-1}(2\mathrm{J} - 1)$, where J is the area under the Lorenz curve for the judged SPD and n is the number of possible events.[1] G_{judge} equals 0 when each category is assigned equal probability and equals 1 when all probability is assigned to a single category. In the marathon example, $n = 5$ and J = 0.668, yielding $G_{judge} = 0.42$.

Given data on the frequencies of each category, we can also construct a Lorenz curve representing the observed distribution in the real-world data. The upper right panel of Figure S9 displays the observed frequencies of finishing times in the marathon. The empirical Lorenz curve is generated by ordering the events from most to least likely according to their observed frequencies while cumulating these empirical probabilities. Letting E denote the area under this curve, the empirical Gini coefficient is given by $G_{empirical} = \frac{n}{n-1}(2\mathrm{E} - 1)$. Finishing times between 3:00 and 3:59 hours were most frequent, with 56% of runners falling into this category. Finishing times over 6:00 hours were least frequent, with only 1% of runners falling into this category. The corresponding empirical Lorenz curve is constructed by successively cumulating the observed event frequencies in decreasing order, yielding E = 0.772 and $G_{empirical} = 0.68$. The quantity $G_{judge} - G_{empirical}$ provides a measure of the extent to which the judge's SPD is more or less concentrated than the observed distribution of events. For this judge, this difference is -0.26, with the negative sign indicating that the judge's SPD is less concentrated than the empirical distribution of finishing times.

Finally, we consider a third Lorenz curve, constructed by cumulating the empirical frequency of events that fall into each category *as they were ordered by the judge*. This calibrated Lorenz curve, displayed in the lower panel of Figure S9, depicts the confidence levels that would be assigned to each of the judge's cumulative categories by a hypothetical judge who knows the empirical distribution perfectly. The associated Gini coefficient is $G_{calibrated} = \frac{n}{n-1}(2\mathrm{C} - 1)$,

[1] Note that we apply an adjustment, $G = \frac{n}{n-1}G^*$, where G^* is the standard Gini coefficient. This is needed in domains in which the population is smaller, such as concentration of market share among companies (e.g., Collins & Preston, 1961) and concentration of crime in particular neighborhoods (Bernasco & Steenbeek, 2017). This correction ensures that the Gini always remains bounded by 0 and 1 (Deltas, 2003).

where C is the area under the calibrated Lorenz curve. Unlike the previous two measures, $G_{calibrated}$ can take negative values, ranging from -1 to 1. The probabilities in our example yield an area of C = 0.468 and $G_{calibrated} = -0.08$. The positive difference $G_{judge} - G_{calibrated} = 0.50$ indicates that the judge's SPD is overprecise.

By defining each of the MCC measures and Gini coefficients formally below, we can see that there is a direct linear correspondence between them.

Definition 1 ($\mathbf{SP_M}$, $\mathbf{EP_M}$, $\mathbf{CP_M}$, $\boldsymbol{G_{judge}}$, $\boldsymbol{G_{empirical}}$, $\boldsymbol{G_{calibrated}}$): Let the probabilities for the judged distribution and the empirical distribution over the intervals be given by $\left(p_1^j, p_2^j, \dots, p_n^j\right)$ and $(p_1^e, p_2^e, \dots, p_n^e)$, respectively, with $\sum_{i=1}^n p_i^j = \sum_{i=1}^n p_i^e = 1$, and $\left(i_1^j, i_2^j, \dots, i_n^j\right)$ and $(i_1^e, i_2^e, \dots, i_n^e)$ be descending orderings which are permutations of (1, *2,…, n*) for each respective distribution such that $p_{i_1^j}^j \geq p_{i_2^j}^j \geq \cdots \geq p_{i_n^j}^j$ and $p_{i_1^e}^e \geq p_{i_2^e}^e \geq \cdots \geq p_{i_n^e}^e$.

The judge's mean cumulative concentration is defined as

$$\begin{aligned}\mathrm{SP_M} &= \left(p_{i_1^j}^j + \left(p_{i_1^j}^j + p_{i_2^j}^j\right) + \cdots + \left(p_{i_1^j}^j + p_{i_2^j}^j + \cdots + p_{i_{n-1}^j}^j\right)\right)/(n-1) \\ &= p_{i_1^j}^j + \frac{n-2}{n-1} p_{i_2^j}^j + \cdots + \frac{1}{n-1} p_{i_{n-1}^j}^j .\end{aligned}$$

The empirical mean cumulative concentration is defined as

$$\begin{aligned}\mathrm{EP_M} &= \left(p_{i_1^e}^e + \left(p_{i_1^e}^e + p_{i_2^e}^e\right) + \cdots + \left(p_{i_1^e}^e + p_{i_2^e}^e + \cdots + p_{i_{n-1}^e}^e\right)\right)/(n-1) \\ &= p_{i_1^e}^e + \tfrac{n-2}{n-1} p_{i_2^e}^e + \cdots + \tfrac{1}{n-1} p_{i_{n-1}^e}^e .\end{aligned}$$

The calibrated mean cumulative concentration is defined as

$$\begin{aligned}\mathrm{CP_M} &= \left(p_{i_1^j}^e + \left(p_{i_1^j}^e + p_{i_2^j}^e\right) + \cdots + \left(p_{i_1^j}^e + p_{i_2^j}^e + \cdots + p_{i_{n-1}^j}^e\right)\right)/(n-1) \\ &= p_{i_1^j}^e + \frac{n-2}{n-1} p_{i_2^j}^e + \cdots + \frac{1}{n-1} p_{i_{n-1}^j}^e .\end{aligned}$$

The Gini coefficient for the judge's subjective distribution is defined as

$$G_{judge} = \tfrac{n}{n-1}(2\mathrm{J} - 1),$$

where J is the area under the Lorenz curve constructed by successively cumulating the judge's probabilities starting at (0,0), increasing linearly to $\left(\frac{1}{n}, p_{i_1^j}^j\right)$, then increasing linearly to

$\left(\frac{2}{n}, p^j_{i^j_1} + p^j_{i^j_2}\right)$, … , increasingly linearly to $\left(\frac{n-1}{n}, p^j_{i^j_1} + p^j_{i^j_2} + \cdots + p^j_{i^j_{n-1}}\right)$, and finally increasingly linearly to (1,1). We can then calculate the area under this curve as

$$\mathrm{J} = \frac{0 + p^j_{i^j_1}}{2n} + \frac{p^j_{i^j_1} + \left(p^j_{i^j_1} + p^j_{i^j_2}\right)}{2n} + \cdots + \frac{\left(p^j_{i^j_1} + p^j_{i^j_2} + \cdots + p^j_{i^j_{n-1}}\right) + 1}{2n}$$
$$= \left(p^j_{i^j_1} + \left(p^j_{i^j_1} + p^j_{i^j_2}\right) + \cdots + \left(p^j_{i^j_1} + p^j_{i^j_2} + \cdots + p^j_{i^j_{n-1}}\right)\right)/2n + 1/2n.$$

The Gini coefficient for the empirical distribution is defined as

$$G_{empirical} = \frac{n}{n-1}(2\mathrm{E} - 1),$$

where E is the area under the Lorenz curve constructed by successively cumulating the empirical probabilities starting at (0,0), increasing linearly to $\left(\frac{1}{n}, p^e_{i^e_1}\right)$, then increasing linearly to $\left(\frac{2}{n}, p^e_{i^e_1} + p^e_{i^e_2}\right)$, … , increasingly linearly to $\left(\frac{n-1}{n}, p^e_{i^e_1} + p^e_{i^e_2} + \cdots + p^e_{i^e_{n-1}}\right)$, and finally increasingly linearly to (1,1). We can calculate the area under this curve as

$$\mathrm{E} = \frac{0 + p^e_{i^e_1}}{2n} + \frac{p^e_{i^e_1} + \left(p^e_{i^e_1} + p^e_{i^e_2}\right)}{2n} + \cdots + \frac{\left(p^e_{i^e_1} + p^e_{i^e_2} + \cdots + p^e_{i^e_{n-1}}\right) + 1}{2n}$$
$$= \left(p^e_{i^e_1} + \left(p^e_{i^e_1} + p^e_{i^e_2}\right) + \cdots + \left(p^e_{i^e_1} + p^e_{i^e_2} + \cdots + p^e_{i^e_{n-1}}\right)\right)/2n + 1/2n.$$

The judge's calibrated Gini coefficient is defined as

$$G_{calibrated} = \frac{n}{n-1}(2C - 1),$$

where C is the area under the Lorenz curve constructed by successively cumulating the empirical probabilities according to the judge's ordering, starting at (0,0), increasing linearly to $\left(\frac{1}{n}, p^e_{i^j_1}\right)$, then increasing linearly to $\left(\frac{2}{n}, p^e_{i^j_1} + p^e_{i^j_2}\right)$, … , increasingly to $\left(\frac{n-1}{n}, p^e_{i^j_1} + p^e_{i^j_2} + \cdots + p^e_{i^j_{n-1}}\right)$, and finally increasingly linearly to (1,1). We can calculate the area under this curve as

$$\mathrm{C} = \frac{0 + p^e_{i^j_1}}{2n} + \frac{p^e_{i^j_1} + \left(p^e_{i^j_1} + p^e_{i^j_2}\right)}{2n} + \cdots + \frac{\left(p^e_{i^j_1} + p^e_{i^j_2} + \cdots + p^e_{i^j_{n-1}}\right) + 1}{2n}$$
$$= \left(p^e_{i^j_1} + \left(p^e_{i^j_1} + p^e_{i^j_2}\right) + \cdots + \left(p^e_{i^j_1} + p^e_{i^j_2} + \cdots + p^e_{i^j_{n-1}}\right)\right)/2n + 1/2n.$$

Note that in case of any ties in ordering for either the judged or actual probabilities, each of these measures is calculated by taking the average over all possible random breaking of the ties. In

practice, this can be estimated by simulating many different random tiebreaks and then averaging the value of the measure over all these instances.

Rearranging terms, we can see that

$$G_{judge} = 2 \times \mathrm{SP_M} - 1,$$

$$G_{empirical} = 2 \times \mathrm{EP_M} - 1, \text{ and}$$

$$G_{calibrated} = 2 \times \mathrm{CP_M} - 1.$$

These definitions of the MCC measures also allow us to formally establish two more results.

Proposition 1: If the judge assigns equal probability mass to all possible categories, then $\mathrm{SP_M} = \mathrm{CP_M} = 1/2$ and $\mathrm{O_M}$ will be zero for any empirical distribution. In other words, totally diffuse beliefs will always be considered well-calibrated (i.e., neither over- nor under-precise).

Proof of Proposition 1: Let the probabilities of the empirical distribution be given by $(p_1^e, p_2^e, \dots, p_n^e)$, with $\sum_{i=1}^{n} p_i^e = 1$. Since the judge assigned equal probability mass to all categories and ties are broken randomly in the calculation of $\mathrm{CP_M}$, we have

$$\mathrm{CP_M} = \frac{1}{n!}\sum_{k=1}^{n!} p_{\sigma_1^k}^e + \frac{n-2}{n-1} p_{\sigma_2^k}^e + \cdots + \frac{1}{n-1} p_{\sigma_{n-1}^k}^e ,$$

where the orderings σ^k denote all $n!$ possible permutations of 1, 2, ..., n. On average, each of the subcomponent probabilities $p_{\sigma^k}^e$ equals $1/n$, since $(p_1^e, p_2^e, \dots, p_n^e)$, are equally represented across each of the summands, meaning that $\mathrm{CP_M} = 1/2$. Next, observe that

$$p_{\sigma_1^k}^j + \frac{n-2}{n-1} p_{\sigma_2^k}^j + \cdots + \frac{1}{n-1} p_{\sigma_{n-1}^k}^j = \frac{1}{n} + \left(\frac{n-2}{n-1}\right)\frac{1}{n} + \cdots + \left(\frac{1}{n-1}\right)\frac{1}{n} = \frac{1}{2}$$

for any ordering of the judge's probabilities, so $\mathrm{SP_M} = 1/2$ and $\mathrm{O_M} = \mathrm{SP_M} - \mathrm{CP_M} = 0$.

Proposition 2: Consider a judge who has uncertainty about which distribution is correct and believes that it could be each one of M different empirical distributions with probabilities w_m, m =1, …, M, satisfying $\sum_{m=1}^{M} w_m = 1$. The judge's predictive distribution should satisfy $\mathrm{SP_M} \leq \mathbf{E}[\mathrm{EP_M}]$, where $\mathbf{E}[\mathrm{EP_M}]$ denotes their expectation of the mean cumulative concentration of the empirical distribution.

Proof of Proposition 2: Denote the empirical probability mass for the categories *1, …, n* of each possible empirical distribution *m* by $(p_1^m, p_2^m, \dots, p_n^m)$, with a corresponding ordering $i_*^m = (i_1^m, i_2^m, \dots, i_n^m)$ such that $p_{i_1^m}^m \geq p_{i_2^m}^m \geq \cdots \geq p_{i_n^m}^m$ and mean cumulative confidence

$$\begin{aligned}\mathrm{EP}_{\mathrm{M}}^m &= \left(p_{i_1^f}^m + \left(p_{i_1^m}^m + p_{i_2^m}^m\right) + \cdots + \left(p_{i_1^m}^m + p_{i_2^m}^m + \cdots + p_{i_{n-1}^m}^m\right)\right)/(n-1) \\ &= p_{i_1^m}^m + \tfrac{n-2}{n-1} p_{i_2^m}^m + \cdots + \tfrac{1}{n-1} p_{i_{n-1}^m}^m .\end{aligned}$$

By the law of total probability, the judge's predictive probability for category *c* is $p_c^j = \sum_{m=1}^M w_m\, p_c^m$, which they should report as their SPD. If $\left(i_1^j, i_2^j, \dots, i_n^j\right)$ is a descending ordering for the judge's distribution, such that $p_{i_1^j}^j \geq p_{i_2^j}^j \geq \cdots \geq p_{i_n^j}^j$, we can rewrite

$$\begin{aligned}\mathrm{SP}_{\mathrm{M}} &= \left(p_{i_1^j}^j + \left(p_{i_1^j}^j + p_{i_2^j}^j\right) + \cdots + \left(p_{i_1^j}^j + p_{i_2^j}^j + \cdots + p_{i_{n-1}^j}^j\right)\right)/(n-1) \\ &= \left(p_{i_1^j}^j + \frac{n-2}{n-1} p_{i_2^j}^j + \cdots + \frac{1}{n-1} p_{i_{n-1}^j}^j\right) \\ &= \sum_{m=1}^M w_m\, p_{i_1^j}^m + \frac{n-2}{n-1} \sum_{m=1}^M w_m\, p_{i_2^j}^m + \cdots + \frac{1}{n-1} \sum_{m=1}^M w_m\, p_{i_{n-1}^j}^m \\ &= \sum_{m=1}^M w_m \left(p_{i_1^j}^m + \frac{n-2}{n-1} p_{i_2^j}^m + \cdots + \frac{1}{n-1} p_{i_{n-1}^j}^m\right)\end{aligned}$$

Importantly, since the calculation of SP_{M} starts by multiplying the largest probability $p_{i_1^j}^j$ by 1, the second-largest probability $p_{i_2^j}^j$ by $\frac{n-1}{n}$, …, and the smallest probability $p_{i_{n-1}^j}^j$ by $\frac{1}{n}$, the calculation of MCC for each of the possible empirical distributions *m* using the descending ordering i_*^m is greater than or equal to the MCC that would be obtained using any other permutation $\tilde{\imath}$ of the n categories as an ordering. Specifically,

$$p_{i_1^j}^m + \tfrac{n-2}{n-1} p_{i_2^j}^m + \cdots + \tfrac{1}{n-1} p_{i_{n-1}^j}^m \leq p_{i_1^m}^m + \tfrac{n-2}{n-1} p_{i_2^m}^m + \cdots + \tfrac{1}{n-1} p_{i_{n-1}^m}^m = \mathrm{EP}_{\mathrm{M}}^m \text{ for all } m.$$

Substituting this inequality into the calculations for SP_{M} yields $\mathrm{SP}_{\mathrm{M}} \leq \sum_{m=1}^M w_m\, \mathrm{EP}_{\mathrm{M}}^m$. ■

In other words, it is normatively appropriate for a judge who has any aleatory uncertainty about which distribution is the correct one to spread out their probability mass so that their concentration is less than or equal to their expectation of the concentration of the epistemic distribution.

S2.2. Comparisons between the MCC measures and other measures of dispersion.

In this subsection, we consider two common alternative calculations that could be used to measure the dispersion of a subjective probability distribution—standard deviation and entropy—and compare them with SP_M on the data from Experiments 1 and 2. The standard deviation of a distribution of a continuous variable which is elicited according to the SPIES procedure (as is the case with our data) can be calculated by fitting a piecewise linear cumulative distribution function (CDF) that passes through the endpoints of each category, ordered from lowest to highest, starting at 0 at the lower bound of the lowest category and ending at 1 at the upper bound of the highest category. The entropy of each distribution can be calculated using the entropy formula for a discrete probability distribution, with the probabilities of each category serving as the discrete set of probabilities. These calculations can be performed for both the judge's SPD and the empirical distribution using the same set of bin endpoints, allowing us to compare the dispersions of subjective and empirical distributions. We use these comparisons to corroborate our hypothesis tests from Experiments 1 and 2 where possible.

What these measures lack, however, is a natural standard by which we can evaluate the *appropriateness* of the dispersion of the judge's distribution. For standard deviation, we can say that a judge with epistemic uncertainty should generally provide an SPD with larger standard deviation than the empirical standard deviation in order to be well-calibrated, but there is no clear normative standard for how exactly how much larger it needs to be. Likewise, for entropy, we can say that the entropy of the judge's SPD should generally be greater than the entropy of the empirical distribution in order to be well-calibrated, but exactly how much greater remains unclear. As a result, we are unable to use these measures to test our main hypotheses relating to overprecision, which limits their usefulness in the main text. However, in presenting the results below, we hope to provide evidence that our conclusions based on our MCC measures agree with the corresponding conclusions that would be obtained using these more familiar measures of dispersion.

We begin by illustrating the strong correlation between SP_M and these alternative measures of dispersion, as evident in Figures S10 and S11. This relationship gives us some reassurance that the overall pattern of results is not driven by our particular choice of how we operationalize the calculation of dispersion, but rather reflects more general regularities in the relationship between judged and empirical dispersions, however they are measured.

Figure S10. The relationship between SP_M and the standard deviation of SPDs in Experiments 1 and 2.

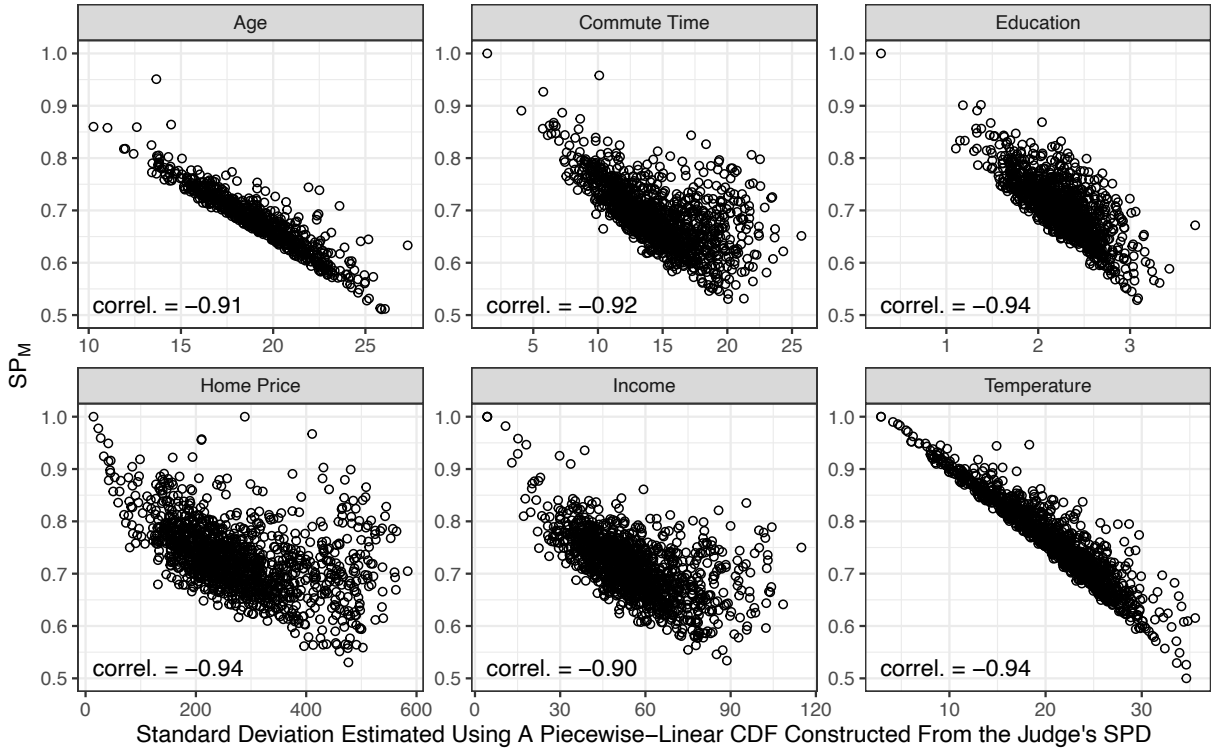
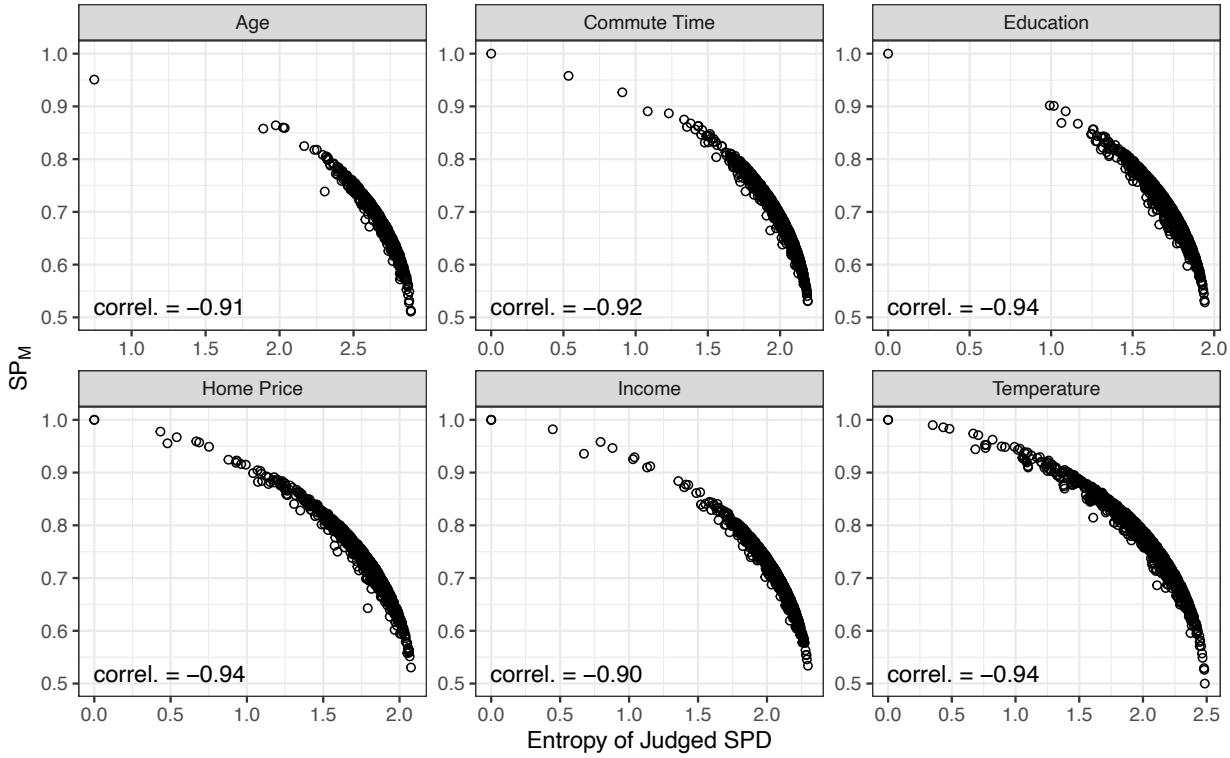


Figure S11. The relationship between SP_M and the entropy of SPDs in Experiments 1 and 2.



To verify this formally, we next revisit hypotheses H1 and H2 from Experiment 1 using standard deviation and entropy (note that H3 cannot be tested using standard deviation or entropy because of the lack of a normative level of dispersion for each individual, as discussed above). As we do for the MCC measures in the main text, we estimate a mixed-model ANOVA with Information (median provided vs. not provided) as a between-subjects variable and Domain as a within-subjects variable. Because the standard deviations are measured in units of the target variable and this scale varies considerably across the domains, we take the logarithm of the ratio of the judged standard deviation and the empirical standard deviation as the dependent variable. This log ratio equals zero when the judged standard deviation matches the empirical standard deviation and is negative (positive) when the judged standard deviation is greater than (less than) the empirical standard deviation. The mean of this log ratio was -0.081 when judges were provided with the median and -0.098 when they did not see the median. Contrary to H1, participants' SPDs were *less dispersed* than the empirical distribution, $F_{1,592} = 100.8$, $p < .001$, $\eta_p^2 = .954$. Furthermore, contrary to H2, participants' SPDs were similarly dispersed as measured by standard deviation regardless of whether or not they saw the median, $F_{1,592} = 0.003$, $p = .957$. Likewise, we analyze dispersion as measured by entropy by estimating a mixed-model ANOVA with the same structure. The mean entropy was 1.89 when judges saw the median and 1.89 when they did not see the median. Contrary to H1, participants' SPDs were *less dispersed* than the empirical distribution, $F_{1,592} = 41.9$, $p < .001$, $\eta_p^2 = .066$. Furthermore, contrary to H2, participants' SPDs were similarly dispersed as measured by entropy regardless of whether or not they saw the median, $F_{1,592} = 0.039$, $p = .844$.

Finally, we revisit hypotheses H1 and H2 in Experiment 2 using standard deviation and entropy (H3 cannot be tested using standard deviation or entropy). As we do for the MCC measures in the main text, we average across the three domains within each round and estimate a 4 (between factor: Information) X 2 (within factor: Round) mixed-model ANOVA. Considering first the standard deviation, we find that the mean of the log ratio of the judged and empirical standard deviation was -0.066 when judges were given the mode and -0.080 when they were not provided with the mode. An ANOVA of this log ratio indicates that, contrary to H1, participants' SPDs were *less dispersed* than the empirical distribution, $F_{1,668} = 100.8$, $p < .001$, $\eta_p^2 = .192$. Participants' SPDs were less concentrated in Round 2 than in Round 1, $F_{1,\,668} = 9.71$, $p = .002$, $\eta_p^2 = .014$. However, H2 was not supported. The log ratio of the judged and empirical standard deviation did

not vary across information conditions, $F_{3,\,668} = 1.43$, $p = 0.23$, $\eta_p^2 = .006$, nor was there an interaction between information condition and round, $F_{3,\,668} = 0.45$, $p = 0.72$, $\eta_p^2 = .002$. Likewise, we analyze dispersion as measured by entropy with a mixed-model ANOVA with the same structure. As before, contrary to H1, participants' SPDs were *less dispersed* than the empirical distribution, $F_{1,668} = 115.6$, $p < .001$, $\eta_p^2 = .148$. Again, H2 was also not supported. The judged entropy did not vary across information conditions, $F_{3,\,668} = 2.91$, $p = 0.088$, $\eta_p^2 = .004$, nor was there an interaction between information condition and round, $F_{3,\,668} = 0.21$, $p = 0.89$, $\eta_p^2 = .001$.

S3. References.

Bernasco W, Steenbeek W (2017) More places than crimes: Implications for evaluating the law of crime concentration at place. *J. Quant. Criminol.* 33(3):451–467.

Collins NR, Preston LE (1961) The size structure of the largest industrial firms, 1909-1958. *Am. Econ. Rev.* 51(5):986–1011.

Deltas G (2003) The small-sample bias of the Gini coefficient: results and implications for empirical research. *Rev. Econ. Stat.* 85(1):226–234.

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* 54(3):306–316.

Gini C (1912) Variabilità e mutabilità. *Repr. Mem. Metodol. Stat. Ed Pizetti E Salvemini T Rome Libr. Eredi Virgilio Veschi*.

Lorenz MO (1905) Methods of measuring the concentration of wealth. *Publ. Am. Stat. Assoc.* 9(70):209–219.

Mamassian P (2008) Overconfidence in an objective anticipatory motor task. *Psychol. Sci.* 19(6):601–606.