Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty

David Tannenbaum, Craig R. Fox, Gülden Ülkümen

To cite this article:
David Tannenbaum, Craig R. Fox, Gülden Ülkümen (2017) Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty. Management Science 63(2):497-518. https://doi.org/10.1287/mnsc.2015.2344

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 2, February 2017, pp. 497–518
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty

David Tannenbaum,[a] **Craig R. Fox,**[b] **Gülden Ülkümen**[c]

[a] David Eccles School of Business, University of Utah, Salt Lake City, Utah 84112; [b] Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90024; [c] Marshall School of Business, University of Southern California, Los Angeles, California 90089

Contact: davetannenbaum@gmail.com (DT); craig.fox@anderson.ucla.edu (CRF); ulkumen@marshall.usc.edu (GÜ)

Received: September 8, 2014
Accepted: July 24, 2015
Published Online in Articles in Advance: March 22, 2016

https://doi.org/10.1287/mnsc.2015.2344

Abstract. People view uncertain events as knowable in principle (*epistemic* uncertainty), as fundamentally random (*aleatory* uncertainty), or as some mixture of the two. We show that people make more extreme probability judgments (i.e., closer to 0 or 1) for events they view as entailing more epistemic uncertainty and less aleatory uncertainty. We demonstrate this pattern in a domain where there is agreement concerning the balance of evidence (pairings of teams according to their seed in a basketball tournament) but individual differences in the perception of the epistemicness/aleatoriness of that domain (Study 1), across a range of domains that vary in their perceived epistemicness/aleatoriness (Study 2), in a single judgment task for which we only vary the degree of randomness with which events are selected (Study 3), and when we prime participants to see events as more epistemic or aleatory (Study 4). Decomposition of accuracy scores suggests that the greater judgment extremity of more epistemic events can manifest itself as a trade-off between enhanced resolution and diminished calibration. We further relate our findings to the hard–easy effect and also show that differences between epistemic and aleatory judgment are amplified when judges have more knowledge concerning relevant events.

History: Accepted by Yuval Rottenstreich, judgment and decision making.
Funding: This work was partially supported by a grant from the National Science Foundation to C. R. Fox and G. Ülkümen [SES-1427469].
Supplemental Material: Data and the supplementary materials are available at https://doi.org/10.1287/mnsc.2015.2344.

Keywords: probability • uncertainty • judgment • accuracy • forecasting • decision analysis

Introduction

Judgment under uncertainty entails two challenges—what to believe and how strongly to hold those beliefs. Determining an appropriate strength of belief is critical for a wide range of decisions by both laypeople and experts. For instance, jurors in U.S. criminal cases not only must determine whether a defendant is more likely guilty than innocent but also must determine whether the defendant is guilty beyond a "reasonable doubt." Physicians are frequently called on to advise their patients not only on what course of treatment to pursue but also of how likely that treatment is to succeed. Consumers confronting a decision to purchase insurance must not only consider whether a future insurance claim is possible but also consider how likely they are to make a claim. Because expectation generally forms the basis of action, formulating appropriate degrees of belief is a necessary component of a rational decision process.

In this paper we focus on *judgment extremity*, the degree to which probabilistic beliefs approach 0 or 1. A well-established literature finds that people are prone to excessive confidence in a wide range of contexts, and that such overconfidence can be both costly and difficult to eliminate (Klayman et al. 1999, Lichtenstein et al. 1982, Moore and Healy 2008). Judgment extremity is the central psychological primitive that defines, relative to empirical frequencies, both overconfidence (judgments that are too extreme) and underconfidence (judgments that are not sufficiently extreme). Moreover, the extremity of one's beliefs determines how much we discriminate between different events and therefore provides a basis for understanding the information contained in a judgment. For example, a person who always estimates a 50% chance that an arbitrarily chosen team will win their baseball game will be well calibrated but not particularly discriminating. Finally, judgment extremity is a critical driver of one's own willingness to act under uncertainty (e.g., Fox and Tversky 1998), and expressions of extremity also strongly influence decisions made by others (e.g., when an eyewitness identifies a potential suspect; Tenney et al. 2007). Thus, understanding the psychological processes that give rise to judgment extremity can shed light on both judgment accuracy and decisions under uncertainty.

We assert that people naturally distinguish two dimensions of uncertainty (Fox and Ülkümen 2011), and this distinction critically influences judgment extremity. First, uncertainty can arise from the inherent unpredictability of random events in the world (as with the uncertainty concerning the outcome of a coin flip); second, uncertainty can arise from awareness of one's deficiencies in knowledge, information, or skills to correctly predict or assess an event that is, in principle, knowable (as with the uncertainty concerning the correct answer to a trivia question). This philosophical distinction between uncertainty of inherently stochastic events (*aleatory* uncertainty) and uncertainty in assessments of what is or will be true (*epistemic* uncertainty) can be traced to the early foundations of probability theory (Hacking 1975), but it has thus far received scant empirical attention as a descriptive feature of judgment under uncertainty.

Across four studies, we find that judgments are more extreme for events viewed as more epistemic and less extreme for events viewed as more aleatory. Prior research suggests that judged probabilities can be modeled as an assessment of the balance of evidence for and against a hypothesis that is mapped onto a number between 0 and 1 (Tversky and Koehler 1994, Rottenstreich and Tversky 1997, Fox 1999). We find that the impact of epistemicness and aleatoriness on judgment extremity can be traced to the mapping of relative evidence onto a judged probability, rather than by perturbing initial impressions of evidence strength.

In the section that follows, we elaborate on the distinction between epistemic and aleatory uncertainty and motivate its connection to judgment extremity. Next, we present a series of empirical tests of our central hypothesis and show how these claims can be embedded within a formal model of judged probability (Tversky and Koehler 1994). In the final section of the paper, we extend our investigation to an analysis of judgment accuracy. We demonstrate that perceptions of epistemic and aleatory uncertainty have opposing effects on distinct components of judgment accuracy—namely, calibration and resolution. We also discuss implications of our findings for improving accuracy in task environments that lead to systematic overconfidence versus underconfidence.

Epistemic vs. Aleatory Judgment Under Uncertainty

Most theories of judgment and decision making construe uncertainty as a unitary construct. For instance, in Bayesian decision theories, subjective probabilities are treated as degrees of belief (e.g., Savage 1954), regardless of their source. Meanwhile, frequentist accounts of probability restrict their attention to situations in which there are stable long-run relative frequencies of classes of events (e.g., von Mises 1957). Fox and Ülkümen (2011) proposed that this historical bifurcation of probability is mirrored by intuitive distinctions that people naturally make between different dimensions of uncertainty. For the purposes of this paper, we distinguish events whose outcomes are viewed as potentially knowable (epistemic uncertainty) from events whose outcomes are viewed as random (aleatory uncertainty). We note that this distinction should be viewed as psychological rather than ontological, and that many judgment tasks are construed as entailing a mixture of these two dimensions. In the current studies, we measure perceptions of epistemic and aleatory uncertainty using a short psychological scale that appears to reliably capture this distinction.

Several lines of research suggest that people naturally distinguish between epistemic and aleatory uncertainty. For instance, 4- to 6-year-old children tend to behave differently when facing chance events yet to occur (in which aleatory uncertainty is presumably salient) versus chance events that have already been resolved but not yet revealed to them (in which epistemic uncertainty is presumably salient; Robinson et al. 2006). Meanwhile, brain imaging studies (Volz et al. 2004, 2005) have found distinct activation patterns when participants learn about events whose outcomes were determined in a rule-based (presumably epistemic-salient) manner compared with a stochastic (presumably aleatory-salient) manner. Furthermore, studies of natural language use suggest that people rely on distinct linguistic expressions to communicate their degree of epistemic and aleatory uncertainty (Ülkümen et al. 2016). In particular, they tend to use words such as "sure" and "confident" when epistemic uncertainty is most salient (e.g., "I am pretty sure that the capital of California is Sacramento"), whereas they tend to use words such as "chance" and "likelihood" when aleatory uncertainty is most salient (e.g., "I think there is a good chance that I'll win this hand of blackjack").

Implications for Judgment Extremity

To see how epistemic and aleatory uncertainty might affect judgment extremity, it is useful to consider a simple generic account of judgment under uncertainty. Once one has identified a target event or hypothesis and its alternatives, one must assess the strength of evidence for each hypothesis and map these impressions onto an explicit expression of belief strength such as a probability judgment (e.g., Tversky and Koehler 1994). Mapping beliefs onto a probability requires one to integrate information concerning the perceived balance of evidence with information concerning its validity or diagnosticity (Griffin and Tversky 1992). For instance, when judging from a political poll how likely it is that a candidate will win an imminent election, a campaign

strategist should consider not only the proportion of respondents favoring his candidate (the strength of the evidence) but also the size, reliability, and representativeness of the poll (the weight of the evidence). Thus, a particular impression of relative evidence strength should be mapped onto a more extreme judgment to the extent that the judge views this impression as a reliable or valid signal, and should be mapped onto a less extreme judgment to the extent that the judge views this impression as unreliable or invalid.

The distinction between epistemic and aleatory uncertainty has an obvious connection to the perceived weight or diagnosticity of evidence. Holding information and the level of knowledge constant, one's impression of relative evidence strength should appear more valid to the extent that the underlying uncertainty is viewed as potentially knowable, predictable, or subject to consensus among experts (epistemic uncertainty) and less valid to the extent that the underlying uncertainty is viewed as inherently random, unpredictable, or variable (aleatory uncertainty). To illustrate, suppose one is predicting which of two players will win a chess match and which of two players will win a hand of poker. Suppose further that one believes the strength imbalance between the two players is the same in both cases (e.g., one player is 25% stronger than his opponent). Assuming that the judge sees the outcome of the chess match as more inherently predictable than the game of poker (based on the relative strength of the two players) and sees a greater role of chance in poker than in chess, it seems apparent that this judge would report a higher probability that the stronger player will prevail in chess than in poker. In short, we predict that the same impressions of relative evidence strength will be mapped onto more or less extreme judgments depending on the perceived nature of the underlying uncertainty.

Naturally, most events entail a mixture of epistemic and aleatory uncertainty, and so the relative impact of these dimensions on judgment extremity may depend on where attention happens to be drawn. For example, consider instances in which two sports teams play each other on multiple occasions in a season. An individual game between the two teams can be viewed as a unique, singular event—occurring on a particular date and time, with a particular lineup of players and coaching strategies—or as an instance drawn from a distribution of roughly exchangeable events. If a judge is asked to consider a matchup between the two teams without an explicit reminder that the game is one of multiple similar matchups, she may be apt to focus on the predictable features of the particular event. Viewed from this perspective, the judge may focus on elements of the matchup that are fundamentally knowable (notably, the relative strengths and weaknesses of each team) when formulating her probability judgment of which team will prevail. However, if the judge is explicitly reminded that the game is one instance of multiple similar matchups, she may be more apt to think about the stochastic nature of wins and losses and consider that outcomes will vary from occasion to occasion even when one team is clearly stronger than its opponent. When viewed in this light, the judge may come to see her impression of relative team strength as a less perfect predictor of game outcomes and consequently report a more conservative probability. Thus, prompting people to think about an event as a member of a class of similar events may promote more conservative judgment than when they naturally consider the same event as a unique case.[1]

As an initial demonstration, we recruited a sample of 75 National Basketball Association (NBA) basketball fans to provide subjective probabilities for three upcoming basketball games,[2] and presented the questions either in a way that might naturally prompt consideration of singular events (highlighting epistemic uncertainty) or in a way designed to prompt consideration of a distribution of similar events (highlighting aleatory uncertainty). For instance, in the Chicago–Detroit matchup, the instructions were as follows.

> *Singular presentation* ($n = 34$):
> The Chicago Bulls will play the Detroit Pistons on March 21st. What is the probability that the Bulls will win?
>
> *Distributional presentation* ($n = 41$):
> The Chicago Bulls will play the Detroit Pistons on February 20th, March 21st, and April 3rd. What is the probability that the Bulls will win on March 21st?

Results of this simple demonstration are provided in Table 1. For all three matchups, we observe greater extremity—judgments that on average are closer to 0 and 1—in the singular presentation that omits distributional cues. This difference in extremity (calculated as the absolute deviation from 1/2) was highly reliable[3] across the three games, $p = 0.003$, confirming our prediction.[4]

Implications for Judgment Accuracy

We have hypothesized that perceptions of epistemic and aleatory uncertainty will affect judgment extremity through the mapping of relative evidence strength onto degrees of belief, rather than by perturbing initial impressions of evidence strength. Several important implications for judgment accuracy follow from this hypothesis. The most straightforward is that although variation in assessments of epistemic/aleatory uncertainty should affect extremity of beliefs, it should not systematically affect participants' ability to correctly identify outcomes (i.e., "hit rates"). This is because if

Table 1. Mean Judgments for Singular vs. Distributional Presentations

	p(Team A wins)		p(Team B wins)		Mean absolute deviation from 1/2	
Team A vs. B	Distributional	Singular	Distributional	Singular	Distributional	Singular
Bulls vs. Pistons	0.63	0.72	0.35	0.32	0.14	0.21
Raptors vs. Hornets	0.63	0.73	0.36	0.35	0.15	0.19
Grizzlies vs. Clippers	0.60	0.61	0.42	0.38	0.09	0.14

initial impressions of evidence strength are unaffected by the perceived nature of uncertainty, then the alternative that is deemed most likely (e.g., which of two teams will win a head-to-head match) will also be unaffected, even as the judged probability of that alternative is amplified toward 1 or dampened toward 1/2.

If perceptions of epistemic and aleatory uncertainty affect judgment extremity, but not hit rates, then this has specific implications for the constituents of judgment accuracy. Overall rates of accuracy can be decomposed into distinct and interpretable components, the best known of which are judgment calibration and resolution (Murphy 1973). *Calibration* measures the extent to which degrees of belief deviate from empirical frequencies—a forecaster is considered well calibrated, for example, if she assigns probabilities of 0.40 to events that occur 40% of the time, assigns probabilities of 0.60 to events that occur 60% of the time, and so forth. Separate from calibration is judgment *resolution*, or the degree to which a forecaster reliably discriminates between different events. Whereas calibration provides a measure of how close a judgment is to the truth, resolution provides a measure of the information contained in a forecast. Our earlier example of someone who always estimates a 50% chance that any given baseball team will win is an instance of someone whose judgments would be well calibrated but lacking in resolution. Thus, increasing judgment extremity while holding hit rates constant should generally increase resolution at the expense of calibration: resolution will tend to improve because participants make fuller use of the probability scale, but calibration will tend to suffer (assuming a general propensity toward overconfidence) because more extreme judgments will increasingly deviate from empirical base rates. We further note that in rare instances where people tend toward underconfidence we would expect increased judgment extremity to improve both resolution and calibration.

Our hypothesis that judgment extremity increases with perceived epistemicness (and decreases with perceived aleatoriness) can also help to explain prior reports of differences in overconfidence across domains. As other researchers have noted (Keren 1991, Wright and Ayton 1987), studies documenting overconfidence have typically relied on general knowledge items such as trivia questions (e.g., Lichtenstein et al. 1982). Note that uncertainty concerning whether one has correctly answered a trivia question will tend to be experienced as purely epistemic (very knowable, not very random). Thus, the present account predicts that judgment extremity, and therefore overconfidence, will tend to be more pronounced for general knowledge questions than for other domains that are seen less epistemic or more aleatory, such as future geopolitical events or sporting matches (Carlson 1993, Fischhoff and Beyth 1975, Howell and Kerkar 1982, Ronis and Yates 1987, Wright 1982, Wright and Wisudha 1982). Indeed, Ronis and Yates (1987) documented greater overconfidence when responding to trivia questions than upcoming professional basketball games, and Wright and Wisudha (1982) documented greater overconfidence for trivia questions than for then-future events. Likewise, in a review of the overconfidence literature, Keren (1991) noted that whenever proper calibration had been identified in prior studies, it was in connection with tasks that involved highly exchangeable events (i.e., those that suggest a natural class of essentially equivalent events), a feature that we surmise promotes distributional thinking and therefore salience of aleatory uncertainty.

Overview of Studies

In this paper we have proposed that attributions of epistemic and aleatory uncertainty influence judgment extremity. We emphasize that these assessments are subjective and can vary across judgment tasks, across individuals assessing the same task, and even within individuals whose impressions of epistemicness or aleatoriness vary with their state of mind. In the studies that follow, we test all of these propositions.

We begin by providing initial evidence that judgment extremity varies systematically with individual differences in perceived epistemicness and aleatoriness (Study 1). We next provide a simple mathematical framework that allows us to formally test our hypothesis about the mapping of evidence onto judged probabilities across different domains (Study 2), within a single judgment domain in which we manipulate relative epistemicness and aleatoriness (Study 3), and in a situation in which we prime participants to view a task as more epistemic or aleatory (Study 4).

Following this exploration of judgment extremity, we examine judgment accuracy across all relevant studies. As predicted, we consistently find that more extreme probability judgments entail a trade-off between different components of judgment accuracy. In particular, perceptions of greater epistemicness are generally associated with increased resolution of probability judgments (i.e., better discrimination) at the expense of decreased calibration (i.e., greater overconfidence). Moreover, we document that the observed pattern of judgment extremity has different implications for judgment accuracy (and therefore corrective strategies) in task environments for which questions are relatively easy versus difficult.

Study 1: Judgment Extremity Increases with Perceived Epistemicness

For Study 1, we recruited a sample of basketball fans and asked participants to provide subjective probabilities for games in the first round of the 2015 National Collegiate Athletic Association (NCAA) men's college basketball tournament. We expected individuals to vary in their beliefs concerning the degree of epistemic and aleatory uncertainty involved in predicting basketball games, and we expected that such differences would covary with the extremity of their judgments. In particular, fans who view basketball games as entailing primarily epistemic uncertainty should provide more extreme judgments than basketball fans who view games as entailing primarily aleatory uncertainty. This task provides a clean first test of our hypothesis, as the tournament is organized around seeded rankings for each team that serve as a natural proxy for consensus estimates of relative team strength. Furthermore, the first round of the tournament provides a number of matchups between teams that vary widely in their degree of parity (e.g., a 1st-seeded team playing a 16th-seeded team, an 8th-seeded team playing a 9th-seeded team), allowing us to examine judgments that should span a wide range of probabilities.

Our sample consisted of 150 college basketball fans (31% female, mean age = 44 years, age range: 21–76 years) who were recruited through an online panel maintained by Qualtrics.com and who were paid a fixed amount for their participation. For this study and all subsequent studies, we determined sample size in advance and terminated data collection before analyzing the results. Before starting the study, participants were asked to report, on seven-point scales, the extent to which they considered themselves fans of college basketball, followed college basketball, and felt knowledgeable about college basketball (e.g., 1 = not at all, 7 = very much so). Only participants who rated themselves at or above the midpoint to all three questions were allowed to proceed to the study. This left us with a sample of fans who expressed familiarity with the judgment domain—our respondents reported watching a median of 3.5 college basketball games per week and a median of 25 total games for the regular season.

After the initial screening questions, participants provided probability judgments for 28 games from the upcoming first round of the NCAA tournament.[5] For each trial, participants were reminded of each team's seeded ranking and judged the probability that a designated team would beat their opponent using a 0%–100% scale. We randomized the order of trials, as well as the team designated as focal for each trial[6] (i.e., whether participants judged $p(A \text{ defeats } B)$ or $p(B \text{ defeats } A)$). To incentivize thoughtful responses, we told participants that some respondents would be selected at random and awarded a bonus of up to $100, in proportion to their accuracy (based on their Brier score; see the supplementary materials for the full task instructions).

Next, participants were presented with three of their earlier games, each randomly sampled from the set of 28 games. For each game, participants rated the degree of epistemic and aleatory uncertainty associated with determining the outcome of the game. This was done using a 10-item epistemic–aleatory rating scale (EARS) that has been developed and validated elsewhere (Fox et al. 2016). The scale prompted participants to rate their agreement with a set of statements that measured feelings of both epistemic uncertainty (e.g., "determining which team will win is something that becomes more predictable with additional knowledge or skills") and aleatory uncertainty (e.g., "determining which team will win is something that has an element of randomness"). For the studies reported here, we reverse-coded the aleatory items and then averaged all responses to form a single "epistemicness" index. Scores on this index take on a value between 1 and 7, with higher numbers indicating a belief that the judgment task entails primarily epistemic uncertainty and lower numbers indicating a belief that the task entails primarily aleatory uncertainty.[7] For Study 1, the Cronbach's α for the EARS scale was 0.70.

Following the disclosure guidelines recommended by Simmons et al. (2011), we provide all materials and measures used for this study, as well as all subsequent studies, in the supplementary materials.

Study 1 Results

We hypothesized that judgments would become increasingly extreme as basketball outcomes were viewed as increasingly epistemic. To test this hypothesis, we estimate the following linear relationship:

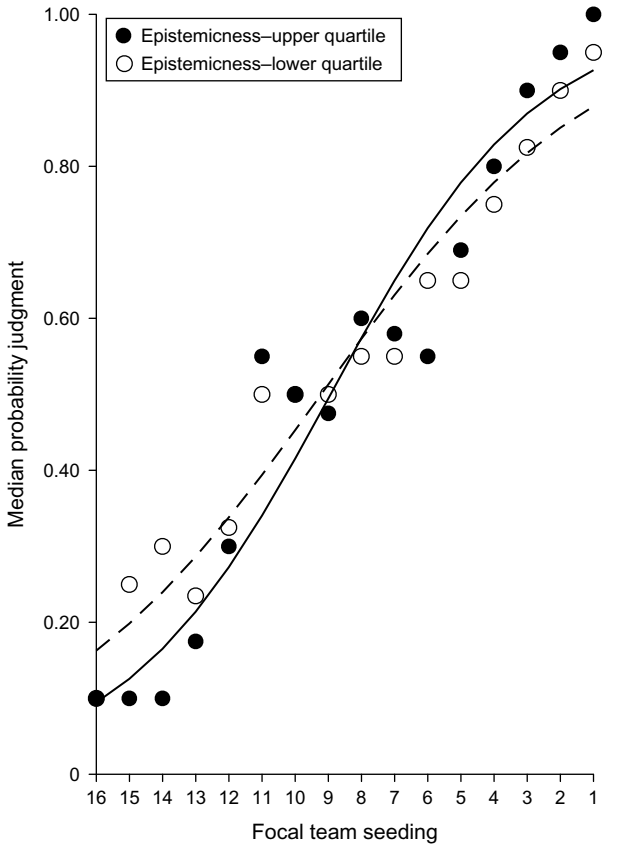
$$Extremity_{ij} = \alpha + \beta_1 Epistemicness_{ij} + U_i + \gamma_j + \epsilon_{ij}, \quad (1)$$

where our dependent variable $Extremity_{ij}$ represents the absolute deviation in judged probability from 1/2 by participant i for basketball game j; responses could take on a value from 0 (a judged probability of 1/2) to 0.50 (a judged probability of either 0 or 1). Our primary predictor of interest, $Epistemicness_{ij}$, represents the epistemicness rating by participant i for game j. We also hold all basketball games fixed in the analysis by including a vector of dummy variables represented by γ_j and model participants as a random effect U_i to account for nonindependence of observations within participants. For this study and all subsequent studies, we conducted analyses using the general linear model described above as well as a fractional logit model that assumes outcomes are bounded between 0 and 1 (Papke and Wooldridge 1996). We find that both models return similar results, so for purposes of simplicity, we report only the results from the linear model.

As predicted, judgment extremity increases with perceived epistemicness; for every one-unit increase in rated epistemicness (on our seven-point index), judgment extremity was expected to increase by 3.7 percentage points[8] ($B = 0.037$, SE = 0.01, $p < 0.001$). This pattern is illustrated in Figure 1, which plots the median judged probabilities for the upper and lower quartiles in rated epistemicness for each of the 16 focal seed rankings. As the figure illustrates, we see more extreme judgments—both higher highs and lower lows—for when participants viewed the NCAA tournament games as more epistemic. Indeed, our results hold both when restricting the analysis to judgments above 1/2 and below 1/2: epistemicness ratings were positively correlated with judged probabilities above 0.50 ($B = 0.042$, SE = 0.01, $p < 0.001$) and negatively correlated with judged probabilities below 0.50 ($B = -0.029$, SE = 0.01, $p = 0.007$). Thus, it appears our results are not driven by a tendency toward greater extremity only when considering probable or improbable events.

We also examined the likelihood of expressing a judgment of complete certainty by dichotomizing responses into certain (i.e., a response of 0 or 1) and uncertain judgments. Using a logit model with the same specification in Equation (1), we again found a greater willingness to express complete certainty as a function of rated epistemicness, with an average marginal effect[9] of 5.5% ($p = 0.022$). Summary statistics of judgment extremity for this study, as well as all subsequent studies, are presented in Table 2.

Figure 1. Study 1: Relationship Between Judgment Extremity and Rated Epistemicness



Notes. The x axis represents the seeded ranking for the target team, and the y axis represents judged probability. For each seeding, median judgments were calculated for the upper and lower quartile of responses in rated epistemicness. Lines represent the best-fitting lines from a fractional response model.

Connecting Judgment Extremity to Evidence Sensitivity

Study 1 demonstrates that judged probabilities were especially extreme for individuals who viewed basketball outcomes as particularly epistemic. In Studies 2–4, we examine how differences in judgment extremity can be attributed to differences in *sensitivity to evidence*—that is, how people map their impressions of evidence strength onto a judged probability. As we discussed in the introduction, we expect that heightened perceptions of epistemic uncertainty will lead to greater sensitivity to differences in evidence strength (i.e., small differences in the strength of evidence between competing hypotheses should translate into more extreme probability judgments). Conversely, heightened perceptions of aleatory uncertainty should lead to diminished evidence sensitivity and relatively regressive judgments, holding the strength of evidence constant. Examining evidence strength can therefore help to explain the judgment extremity effect we observed in Study 1.

There is another important reason for investigating evidence sensitivity. Doing so allows us to examine judgment extremity across domains while controlling for parity in the strength of hypotheses drawn from each domain. To illustrate this point, suppose we

Table 2. Epistemicness Ratings and Judgment Extremity in Studies 1–4

		Judgment extremity			
	Epistemicness M (SD)	MAD from $p = 0.50$	Median $p > 0.50$	Median $p < 0.50$	Proportion $p = 0$ or 1
Study 1					
Fourth quartile	4.97 (0.45)	0.29	0.85	0.14	0.16
Third quartile	4.35 (0.10)	0.24	0.75	0.23	0.08
Second quartile	3.87 (0.19)	0.23	0.70	0.20	0.03
First quartile	2.79 (0.48)	0.18	0.70	0.30	0.00
Study 2					
Geography	6.45 (1.00)	0.29	0.80	0.10	0.28
Oceans	6.35 (1.13)	0.36	0.98	0.10	0.43
Population	5.96 (1.21)	0.33	0.90	0.10	0.15
Crime	4.52 (1.56)	0.31	0.80	0.20	0.13
Housing	3.96 (1.53)	0.20	0.70	0.30	0.02
Temperature	3.21 (1.19)	0.21	0.75	0.28	0.01
Rain	3.11 (1.27)	0.15	0.65	0.30	0.01
Movies	3.05 (1.45)	0.23	0.80	0.25	0.08
Politics	2.95 (1.15)	0.16	0.60	0.35	0.05
Baseball	2.49 (1.26)	0.11	0.65	0.30	0.02
Football	2.41 (1.03)	0.11	0.65	0.30	0.00
Soccer	2.40 (1.17)	0.10	0.65	0.40	0.02
Study 3					
Yearlong average task	4.96 (1.06)	0.30	0.80	0.10	0.20
Arbitrary day task	4.35 (1.11)	0.25	0.80	0.20	0.07
Study 4					
Pattern detection	4.38 (0.87)	0.21	0.76	0.25	0.02
Random prediction	4.28 (0.84)	0.18	0.70	0.30	0.01

Notes. MAD = mean absolute deviation. Average probability estimates for each question per study are reported in the supplementary materials.

asked participants to judge the probability that various basketball teams and football teams will win their upcoming games. More extreme probabilities for football than basketball could reflect differences in beliefs about the degree of epistemic and aleatory uncertainty underlying each sport, but such differences in judgment extremity could also simply reflect a belief that the selection of football teams was more imbalanced than the selection of basketball teams. Controlling for explicit ratings of evidence strength allows us to remove this potential confound and provides a common metric by which we can compare judgments across domains.[10]

Sensitivity to evidence strength can be formalized using support theory (Tversky and Koehler 1994, Rottenstreich and Tversky 1997). In support theory, probabilities are attached to *hypotheses*, or descriptions of events.[11] Each hypothesis A is associated with a nonnegative support value, $s(A)$. Support values can be thought of as impressions of the strength of evidence favoring a particular hypothesis—evoked by judgment heuristics, explicit arguments, or other sources. According to support theory, judged probability is a function of the support for a focal hypothesis relative to the support for its complement. That is, the probability $p(A, B)$ that the focal hypothesis A rather than the complementary hypothesis B obtains is given by

$$p(A, B) = \frac{s(A)}{s(A) + s(B)}. \tag{2}$$

Support is a latent construct that can only be inferred from probability judgments. However, it is possible to link hypothetical support, $s(\cdot)$, to a raw measure of evidence strength, $\hat{s}(\cdot)$. This is accomplished by relying on two modest assumptions that have been empirically validated in prior research (Fox 1999, Koehler 1996, Tversky and Koehler 1994). First, direct assessments of evidence strength and support values (derived from judged probabilities) are monotonically related: $\hat{s}(A) \geq \hat{s}(B)$ iff $s(A) \geq s(B)$. Note that this condition implies that $\hat{s}(A) \geq \hat{s}(B)$ iff $p(A, B) \geq 1/2$. For instance, if $\hat{s}(\cdot)$ refers to the strength of basketball teams and $p(A, B)$ is the judged probability that team A beats team B, then this assumption merely implies that a judge will rate team A at least as strong as team B if and only if she judges the probability that team A will beat team B to be at least 1/2. Second, corresponding strength and support ratios are monotonically related: $\hat{s}(A)/\hat{s}(B) \geq \hat{s}(C)/\hat{s}(D)$ iff $s(A)/s(B) \geq s(C)/s(D)$. This assumption implies that the higher the ratio of judged strength between the focal and alternative hypotheses,

the higher the judged probability of the focal hypothesis relative to the alternative hypothesis. For instance, the relative strength of team A to team B should be at least as high as the relative strength of team C to team D if and only if the judged probability of team A beating team B is at least as high as the judged probability of team C beating team D.[12]

If these two conditions hold, and support values are defined on, say, the unit interval, then it can be shown that there exists a scaling constant $k > 0$ such that measures of strength are related to support by a power transformation of the form $s(A) = \hat{s}(A)^k$ (see Theorem 2 of Tversky and Koehler 1994). Intuitively, one can interpret the scaling constant k as an index of an individuals' sensitivity to differences in evidence strength when judging probability. This interpretation can be seen more easily by converting probabilities into odds. Using Equation (1), assuming all probabilities are positive, and defining $R(A,B)$ as the odds that A rather than B obtains, we get

$$R(A,B) \equiv \frac{p(A,B)}{1-p(A,B)} = \frac{s(A)}{s(B)} = \left[\frac{\hat{s}(A)}{\hat{s}(B)}\right]^k. \qquad (3)$$

We see from this equation that as k approaches 0, $R(A,B)$ approaches 1 and probabilities converge toward the ignorance prior of 1/2. When k is equal to 1, we see a linear mapping between the balance of evidence strength $\hat{s}(A)/[\hat{s}(A)+\hat{s}(B)]$ and judged probability $p(A,B)$. As k increases above 1, subjective probability will increasingly diverge to 0 or 1 as differences in evidence strength emerge (see Figure 2). Thus, our hypothesis implies that k should increase when tasks are viewed as more epistemic,[13] and decrease when tasks are viewed as more aleatory.

This formulation also allows us to easily recover k (i.e., evidence sensitivity) from raw strength ratings and judged probabilities. To do so, we simply take the logarithm of both sides of Equation (2):

$$\ln R(A,B) = k \ln \frac{\hat{s}(A)}{\hat{s}(B)}. \qquad (4)$$

Thus, using Equation (4), we can empirically estimate sensitivity to evidence strength by regressing log odds (derived from judged probabilities) onto log strength ratios, with the coefficient from the log strength ratio providing an estimate of k. In the studies that follow, we use this approach when probing for differences in sensitivity to evidence strength.

Study 2: Differences in Evidence Sensitivity Across Domains

In Study 2 we examined evidence sensitivity across a wide variety of domains, with the prediction that across-domain differences in evidence sensitivity would be positively correlated with across-domain differences in judged epistemicness. We recruited a sample of 205 participants from Amazon's Mechanical Turk (MTurk) labor market and paid them a fixed amount in return for their participation[14] (56% female, mean age = 33 years, age range: 18–80 years). One participant reported using outside sources (e.g., *Wikipedia*) to complete the task and was dropped from the analysis. Participants first provided probability judgments to 6 questions that were randomly sampled from a pool of 12 questions, listed in Table 3, with each question drawn from a different topic domain. The order of trials was randomized, and for each trial we counterbalanced which of the two targets was designated as focal.

Next, participants provided strength ratings for the two targets in each of their six previous estimates (following Tversky and Koehler 1994). For each question, participants were asked to assign a strength rating of 100 to the stronger of the two targets and then to scale the other target in proportion. For example, the strength rating procedure for the football question was as follows:

> Consider the Arizona Cardinals and the San Francisco 49ers. First, choose the football team you believe is the stronger of the two teams, and set that team's strength rating to 100. Assign the other team a strength rating in proportion to the first team. For example, if you believe that a given team is half as strong as the first team (the one you gave 100), give that team a strength rating of 50.

Figure 2. Examples of Sensitivity to Evidence Strength (k)

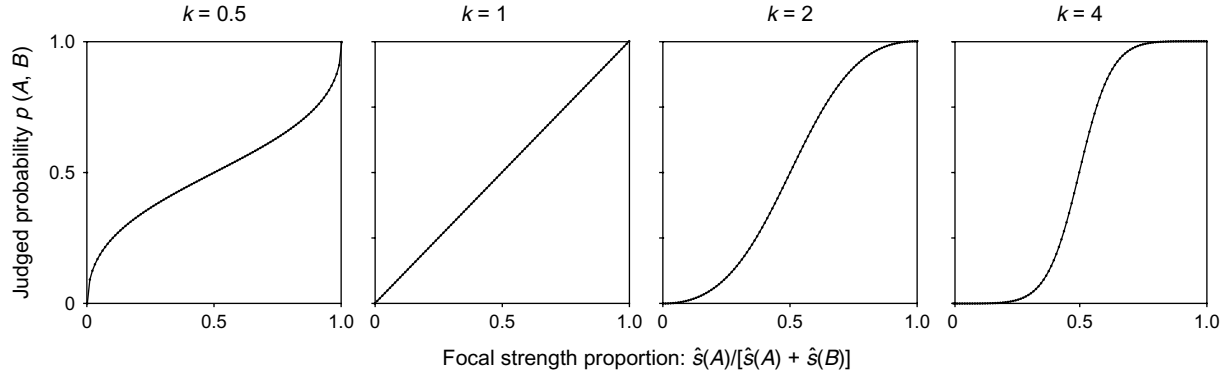


Table 3. Study 2A Questions

Domain	Question
Rain	Consider the weather in Chicago and Minneapolis. What is the probability that there will be more rainy days next May in Chicago than in Minneapolis?
Temperature	Consider the weather in Portland and Pittsburgh. What is the probability that the daytime high temperature next June 1st will be higher in Portland than in Pittsburgh?
Politics	Assume that Barack Obama will face Mitt Romney in the 2012 presidential election. What is the probability that Barack Obama will beat Mitt Romney?
Football	The San Francisco 49ers will play the Arizona Cardinals on October 29th. What is the probability that the San Francisco 49ers will beat the Arizona Cardinals?
Baseball	The Chicago Cubs will play the Los Angeles Dodgers on August 3rd. What is the probability that the Chicago Cubs will beat the Los Angeles Dodgers?
Movies	Consider two upcoming summer movies, *The Amazing Spider-Man* and *The Dark Knight Rises*. What is the probability that *The Amazing Spider-Man* will gross more money on its opening weekend than *The Dark Night Rises*?
Housing	Consider housing prices in Nashville and Atlanta. What is the probability that a randomly selected house in Nashville will be more expensive than a randomly selected house in Atlanta?
Crime	Consider crime rates in Detroit and Columbus. What is the probability that the number of violent crimes per capita this year will be higher in Detroit than in Columbus?
Geography	Consider the geographic size (in square miles) of Nevada and Wyoming. What is the probability that Nevada is larger than Wyoming?
Population	Consider the urban population of Istanbul, Turkey and Shanghai, China. What is the probability that Istanbul has a larger urban population than Shanghai?
Soccer	Suppose the Italian national soccer team plays Germany this summer in the European Cup. What is the probability Italy will beat Germany?
Oceans	Consider the size (in square miles) of the Atlantic Ocean and Indian Ocean. What is the probability that the Atlantic Ocean is larger than the Indian Ocean?

Finally, participants were again shown each of the six events they had previously assessed and rated each event using an abridged four-item epistemicness scale (Cronbach's α ranged from 0.60 to 0.87 across domains, with an average score of 0.75; see the supplementary materials for scale items).

Data Exclusions

For Study 2, as well as all subsequent studies, we excluded a small number of responses where estimated probabilities fell outside of the 0–100 range or where participants provided a strength rating of 0 to either the focal or alternative target.[15] Such responses suggest a misunderstanding of the task scale and are not directly interpretable (i.e., cannot be analyzed without retransforming the data). We also excluded participants whose judgments revealed negative evidence sensitivity, which mostly likely implies inattentive responding—taking negative k estimates seriously would mean that participants find hypotheses with less relative evidence strength *more* probable than hypotheses with greater relative evidence strength. These exclusion rules required us to drop no more than 9% of participants per study (18 participants in Study 2, 12 participants in Study 3, and 6 participants in Study 4). For all studies, retaining these "problematic" participants in the analysis does not qualitatively change the results.

Analysis Strategy

For Studies 2–4, we test for judgment extremity in a manner similar to Study 1 by estimating the following relationship:

$$Extremity_{ijk} = \alpha + \beta_1 Treatment_k + U_i + \gamma_j + \epsilon_{ijk}, \quad (5)$$

where $Extremity_{ijk}$ represents the absolute deviation in judged probability from 1/2 by participant i for question j in treatment k. $Treatment_k$ represents the treatment variable of interest. In Study 2, this term represents a set of indicator variables for each judgment domain; in Studies 3 and 4, the treatment variable represents an indicator variable for the specific judgment task. We always include participant random effects, denoted by U_i, to account for nonindependence of observations within participants. When appropriate,[16] we also control for variation in question items by modeling them as fixed effects, denoted by γ_j.

In Studies 2–4, we also test for differences in sensitivity to evidence strength. As discussed earlier, an analysis of evidence sensitivity allows us to more rigorously account for the nature of events in our sample and their distribution (e.g., an analysis of evidence sensitivity controls for domain-level differences in parity of strength ratings). As suggested by Equation (4), we estimate evidence sensitivity by first transforming judged probabilities into log odds[17] (i.e., $\ln[p(A,B)/(1-p(A,B))]$ for judgments $p(A,B)$), and we regress log odds onto the log strength ratios for the hypotheses under consideration (i.e., $\ln[\hat{s}(A)/\hat{s}(B)]$):

$$LogOdds = \alpha + \beta_1 LogStrength + \epsilon. \quad (6)$$

In Equation (6), the observed coefficient for *LogStrength* can be interpreted as an index of evidence sensitivity, with higher numbers indicating greater sensitivity.[18] Because in all studies we are interested in examining differences in evidence sensitivity across our treatment

variable(s) of interest, this requires we estimate the following relationship:

$$LogOdds_{ijk} = \alpha + \beta_1 LogStrength_{ij} + \beta_2 Treatment_k + \beta_3 LogStrength_{ij} \times Treatment_k + U_i + \gamma_j + \epsilon_{ijk}, \quad (7)$$

where $LogOdds_{ijk}$ represents the log odds by participant i for question j in treatment k. $LogStrength_{ij}$ represents the log strength ratio by participant i for question j. $Treatment_k$ is a vector of indicator variables representing the treatment variable of interest. We are interested in examining differences in evidence sensitivity across treatment conditions, so we include a vector of interaction terms, denoted by $LogStrength_{ij} \times Treatment_k$, that model the change in the strength ratio coefficients as a function of our treatment variable(s). Thus, estimating the model above allows us to recover estimates of evidence sensitivity for each treatment condition by calculating the slope of *LogStrength* conditional on that treatment. Again, we include participant random effects and question fixed effects in the analysis.

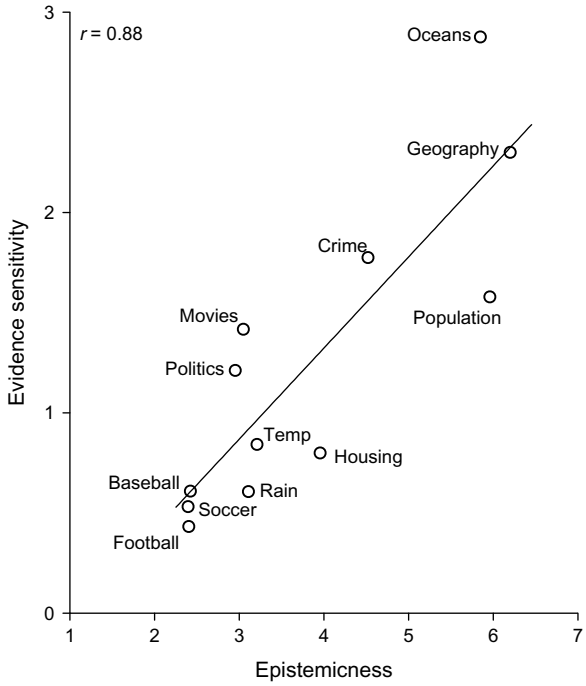
Study 2 Results

Table 2 lists the average epistemicness rating by domain, along with indices of judgment extremity. As the table clearly shows, domains higher in epistemicness also tended to exhibit greater judgment extremity. For each domain, we calculated the mean absolute deviation from 1/2 (the second data column of Table 2), and we correlated these values with corresponding epistemicness ratings (the first data column of Table 2). As expected, the correlation was positive and substantial ($r = 0.91$, $p < 0.001$). We obtain similar results when restricting the analysis to judgments above 0.50, below 0.50, or of 0 and 1 (p-values less than0.001).

Next, we calculated and recovered estimates of evidence sensitivity separately for each domain using the specification detailed in Equation (7). We then correlated these estimates with each domain's mean epistemicness rating. Figure 3, which plots this relationship, indicates that sensitivity to differences in evidence strength was generally higher for domains entailing greater epistemic uncertainty ($r = 0.88$, $p < 0.001$). Using the predicted point estimates from the model, we would expect to see a 4.3-fold increase in evidence sensitivity when going from the domain lowest in epistemicness to the domain highest in epistemicness (i.e., a larger effect than going from the second panel in Figure 2 to the fourth panel).

The above analysis examined the correspondence between rated epistemicness and judgments across domains; the design of Study 2 also allowed us to examine this relationship within individuals. For judgment extremity, we calculated the rank-order correlation between each participants' absolute deviation

Figure 3. Study 2: Relationship Between Evidence Sensitivity (k) and Rated Epistemicness



from 1/2 (for judged probability) and the corresponding epistemicness ratings. Similar to before, we predict a positive correlation between rated epistemicness and judgment extremity, and we observe that 81% of our participants exhibited a positive relationship between their judgment extremity and epistemicness ratings ($p < 0.001$ by a sign test), with a median correlation of $\rho = 0.52$. We also conducted a similar analysis for evidence sensitivity by calculating an analytic, rather than estimated, measure of evidence sensitivity for each observation[19] (i.e., dividing $LogOdds_{ij}$ by $LogStrength_{ij}$ for question j by participant i), and then we computed the rank-order correlation between each participants' evidence sensitivity and epistemicness ratings. Here, we observe even larger effects than those found for judgment extremity: 94% of our participants exhibited a positive relationship between evidence sensitivity and epistemicness ratings ($p < 0.001$ by a sign test), with a median correlation of $\rho = 0.77$.

The results of Study 2 suggest that across-domain differences in evidence sensitivity vary systematically with differences in perceived epistemicness. We note, however, that these differences were estimated from a single question per domain, and so it is possible that we happened to sample idiosyncratic questions from each domain that gave rise to our results. In the supplementary materials, we report the findings from a follow-up study (Study 2S) that focused on a smaller number of domains but more exhaustively sampled target events within each domain. We selected three

judgment domains expected to span the range of perceived epistemicness—geography questions, weather estimates, and upcoming NBA basketball games—and participants provided probability judgments to 16 questions per domain. Consistent with the results of Study 2, we found that both judgment extremity and evidence sensitivity followed the same rank ordering as epistemicness ratings across domains (using a variety of estimation techniques). Thus, the pattern we observe in Study 2 is robust to a more thorough sampling of stimulus questions.

Study 3: Manipulating Diagnostic Value of Evidence Strength

Study 2 demonstrates that domains entailing relatively greater epistemic uncertainty are associated with greater evidence sensitivity and, consequently, greater judgment extremity. One limitation of this study is that different domains require different measures of evidence strength, and it is therefore unclear to what extent the (unobserved) measurement error associated with the elicitation of strength ratings accounts for observed differences in evidence sensitivity. For instance, consider the strength rating measure we used when participants judged the probability that one football team would beat another—namely, the relative overall strength of each team. Suppose that we had instead asked about the relative strength of each *coaching staff*. In this case we surely would have recovered lower values of the k parameter. It is possible that the raw measures of evidence strength we selected for more epistemic domains in Study 2 were, for whatever reason, more appropriate proxies of hypothetical support.[20] Thus, it would be desirable to replicate our analysis for events that are matched in their natural measure of strength but for which we experimentally manipulate epistemicness of the criterion judgment. Such a test would allow us to more carefully examine whether individuals' perceptions of epistemicness across matched domains predict differences in their sensitivity to evidence and extremity of judgment.

To that end, in Study 3 we asked participants on each trial to estimate the probability that one of two U.S. cities had a higher daytime high temperature. In two separate blocks of trials, participants compared cities according to (a) their average temperature from the previous year and (b) an arbitrarily selected day over the same time interval. Naturally, global impressions of evidence strength—in this case, that one city is "warmer" than another—should be more diagnostic of yearlong averages than of single days, since there is greater fluctuation in temperatures over individual days than over an average of a collection of days. For this reason, yearlong average questions should be seen by most participants as more epistemic than single-day questions about the same pairs of cities, and we should generally observe more extreme judged probabilities for yearlong average questions than for single-day questions. More interesting is the question of whether individual differences in perceived epistemicness across the two tasks accounts for corresponding differences in judgment extremity and evidence sensitivity.[21]

Study 3 Methods

We recruited a sample of 199 participants from MTurk who were paid a fixed amount in return for their participation (52% female, mean age = 37 years, age range: 19–70 years). One participant was removed for reporting that he or she used external sources while completing the survey.

All participants completed two blocks of probability judgments, with blocks presented in a random order. For trials in the *yearlong average* block, participants were asked to estimate the probability that one of two U.S. cities had a higher average temperature in the previous year. For trials in the *arbitrary day* block, participants were asked to estimate the probability that one of two cities had a higher temperature on an arbitrarily selected day from the previous year.[22] Table 4 provides sample questions. Each block consisted of 15 trials (by forming all pairwise comparisons between six cities) that were presented in a random order. For each trial, the city designated as focal was counterbalanced between participants but remained fixed within participants across the two blocks. Upon completing the 15 trials within a given block, participants rated the task epistemicness of three randomly selected trials using a 10-item EARS measure similar to that used in Study 1. After responding to both judgment blocks, participants provided strength ratings for the six cities in a manner similar to Study 2.

Study 3 Results

As expected, the yearlong average task was rated on average as entailing greater epistemicness than the arbitrary day task (means were 4.96 and 4.35, respectively; $p < 0.001$). Analyzing the data within participants, 69% of our respondents rated the former as higher in epistemicness than the latter ($p < 0.001$ by a sign test).

Consistent with our hypothesis, we also observed greater judgment extremity in the yearlong average

Table 4. Study 3 Sample Questions

Task format	Sample question
Yearlong average	What is the probability that the average temperature last year was higher in Anchorage than in Indianapolis?
Arbitrary day	What is the probability that the temperature of an arbitrarily selected day from last year was higher in Anchorage than in Indianapolis?

task than in the arbitrary day task (see Table 2); the mean absolute deviation was 4.8 percentage points higher when judging yearlong averages than arbitrarily selected days ($B = 0.048$, SE = 0.006, $p < 0.001$). Participants also displayed greater judgment extremity across the two tasks when restricting the analysis to judgments above 0.50 or below 0.50, or when dichotomizing responses into certain versus uncertain judgments (all p-values less than 0.001). Analyzing the data within participants, 75% of individual respondents displayed greater mean absolute deviation in their judgments of yearlong averages than arbitrarily selected days ($p < 0.001$ by a sign test).

We estimated average evidence sensitivity for the two tasks in a manner similar to what we used in Study 2. Consistent with our judgment extremity results, participants displayed greater evidence sensitivity when responding to yearlong averages than arbitrarily selected days[23] (estimated k values were 2.17 and 1.53, respectively; $B = 0.63$, SE = 0.05, $p < 0.001$). We also examined within-participant differences in evidence sensitivity by calculating each participant's evidence sensitivity score for the two tasks and found that 69% of participants displayed greater evidence sensitivity in the yearlong average task than in the arbitrary day task ($p < 0.001$ by a sign test).

Recall that the main aim of Study 3 was to identify whether variation in impressions of relative epistemicness across the two tasks explained concomitant shifts in evidence sensitivity. At the trial level, our prediction would imply a positive interaction effect between strength ratings and perceived epistemicness—the slope on strength ratings, which represents our estimate of evidence sensitivity, should increase as perceived epistemicness increases. Accordingly, we regressed log odds onto log strength ratios, epistemicness ratings, and the interaction between the two. Accordingly, we regressed log odds onto log strength ratios, epistemicness ratings, and interaction between the two (i.e., similar to the model in Equation (7), but with epistemicness ratings replacing the treatment variable[24]). As expected, we found a reliable and positive interaction effect ($B = 0.23$, SE = 0.06, $p < 0.001$). At the participant level, we examined this by first calculating, for each respondent, the difference in the respondent's degree of evidence sensitivity between the two tasks ($\Delta_k = k_{average} - k_{arbitrary\ day}$) as well as the difference in epistemicness ratings between the two tasks ($\Delta_{epistemicness}$). Thus, we would predict a positive correlation between Δ_k and $\Delta_{epistemicness}$—those who show the largest shifts in rated epistemicness across tasks should also show the largest shifts in evidence sensitivity. As predicted, we observe a positive and significant correlation between the two difference scores ($r = 0.24$, $p < 0.001$). Likewise, a nonparametric analysis reveals that for 63% of participants, evidence sensitivity and perceived epistemicness were rank ordered identically across the two tasks (i.e., the signs of Δ_k and $\Delta_{epistemicness}$ coincided; $p < 0.001$ by a binomial test).

Study 4: Priming Epistemic and Aleatory Uncertainty

In Study 3, we varied a dimension of the judgment task that we expected to influence perceived epistemicness and therefore extremity of judgment. In Study 4, we investigate an even more subtle manifestation of this phenomenon: whether we can prime people to see a fixed event as more or less epistemic and therefore make more or less extreme probability judgments. Such a demonstration would provide an even stronger test of the causal relationship between perceived epistemicness and evidence sensitivity.

To manipulate participants' predisposition to see events in the world as more epistemic or more aleatory, we asked them to perform a simple binary prediction task with an unknown distribution. In these "two-armed bandit" environments, there is a well-documented tendency for an individual's choice proportions to match the relative frequencies with which each option delivers a favorable outcome (i.e., probability matching; Herrnstein 1997). Although this behavior is commonly viewed as suboptimal (because choosing the higher expected value option on every trial will maximize earnings), recent research has suggested that the switching behavior inherent to probability matching may reflect an effort to discern underlying patterns in a task that is seen as not entirely random (Gaissmaier and Schooler 2008, Goodnow 1955, Unturbe and Corominas 2007, Wolford et al. 2004). Accordingly, we varied the task instructions to either promote pattern seeking (thereby making epistemic uncertainty salient) or promote thinking about the relative frequencies of stochastic events (thereby making aleatory uncertainty salient). Our purpose was to see whether perceptions of epistemicness versus aleatoriness on the two-armed bandit task would carry over to a second, ostensibly unrelated task, and if we would observe concomitant shifts in judgment extremity and evidence sensitivity.

Study 4 Methods

We recruited 100 students from a subject pool at the University of California, Los Angles. Each student was paid a fixed amount for participating (82% female, mean age = 20 years, age range: 16–58 years).

The study consisted of four phases. In the first phase, participants completed a binary prediction task where, for each trial, they predicted whether an X or an O would appear next on the screen. This task served as our experimental prime, and our key manipulation was to vary how this first phase of the study was described to participants. In the *pattern detection*

condition, participants were introduced to a "pattern recognition task" and were given the following instructions:

> On each trial, you will try to predict which of two events, X or O, will occur next. The sequence of Xs and Os has been set in advance, and your task is to figure out this pattern.

In the *random prediction* condition, participants were introduced to a "guessing task" and were given the following instructions:

> On each trial, you will try to guess which of two events, X or O, will occur next. The order of Xs and Os will be randomly generated by a computer program, and your task is to guess which outcome will appear next.

After 10 practice trials, all participants completed the same 168 trials divided into two blocks of 84 trials. Because half of participants thought there was a pattern and half did not, we presented half of the trials with a pattern and half without; in one block participants viewed trials that were generated randomly, whereas in the other block trials followed a fixed 12-digit pattern (e.g., XXOXOXXXOOXX; see Gaissmaier and Schooler 2008, for a similar design). The underlying proportion of X's and O's was the same in both blocks, with a 2:1 ratio for the more common letter. The letter designated as more common, as well as the order of the two blocks was counterbalanced across participants. Participants received feedback about the accuracy of their prediction after each trial. To incentivize thoughtful responding, we notified participants that the most accurate respondent would receive a bonus payment of $25. Performance on this task did not vary systematically by either the priming instructions or the ordering of the two trial blocks.[25]

In the second phase of the study, participants provided 28 probability judgments to upcoming weather-related events in eight U.S. cities (which served as the primary dependent variable). For each trial, participants were presented with two cities (sampled from a pool of eight), with one city designated as focal. Participants indicated the probability that the focal city would have a higher daytime high temperature on the following July 1. The order of these trials was randomized, and the city designated as focal was counterbalanced across participants.

In the third phase, participants provided strength ratings (in terms of each city's relative "warmth") for the eight cities, using the same procedure as before. In the final phase of the study, participants were presented with three randomly selected trials from the second phase, and they rated each question on the 10-item EARS used in previous studies (average Cronbach's $\alpha = 0.70$).

Study 4 Results

As predicted, probability judgments were more extreme when participants were primed with pattern detection than random prediction (see Table 2). Using mean absolute deviation from 1/2, judgments were on average 2.9 percentage points more extreme for the pattern detection task than the random prediction task ($B = 0.029$, SE = 0.016, $p = 0.035$). We also observed greater judgment extremity when restricting the analysis to judgments above 0.50 ($B = 0.035$, SE = 0.015, $p = 0.01$) or below 0.50 ($B = -0.035$, SE = 0.017, $p = 0.021$), and we also observed a directional but nonsignificant difference when dichotomizing responses into certain versus uncertain judgments ($p = 0.11$). This last null result is likely because participants reported a complete certainty judgment in only 1% of all trials.

Most important, we observed greater sensitivity to evidence strength in the pattern detection task than in the random prediction task. Calculating evidence sensitivity in a manner similar to previous studies, we find greater evidence sensitivity when primed with pattern detection than random prediction[26] (estimated k values were 1.44 versus 0.97, $B = 0.47$, SE = 0.23, $p = 0.021$).

As a manipulation check, we examined average epistemicness scores for each task. Questions were viewed as entailing more epistemic uncertainty when participants were prompted to seek patterns than when they were prompted to guess, although this difference was not statistically significant (means were 4.39 and 4.27, respectively; $p = 0.22$).[27]

Finally, we examined the relationship between epistemicness and sensitivity to evidence strength. This was done at the trial level by probing for a positive interaction between strength ratings and perceived epistemicness—the coefficient of the log strength ratio, which is an estimate of evidence sensitivity, should increase as perceived epistemicness increases. We regressed log odds onto log strength ratios, epistemicness ratings, and interaction between the two in a manner similar to Study 3. As expected, we found a positive interaction term ($B = 0.16$, SE = 0.10, $p = 0.053$). Based on the regression coefficients, k would be expected to increase from 0.93 to 1.20 when going from one standard deviation below to one standard deviation above the mean in perceived epistemicness.

General Discussion

The current research provides strong evidence that judgment is more extreme under uncertainty that is perceived to be more epistemic (and less aleatory). We observed this pattern among basketball fans who differed in their perceptions of the epistemicness of college basketball games (Study 1), across judgment domains that differ markedly in their perceived epistemicness (Study 2), in a judgment domain for which we manipulated the degree of randomness with which

events were selected (Study 3), and when participants were experimentally primed to focus on epistemic or aleatory uncertainty (Study 4). These results suggest that lay intuitions about the nature of uncertainty may have downstream implications for judgment and choice. In what follows, we discuss theoretical extensions and implications.

Epistemicness and Judgment Accuracy

As outlined in the introduction, judgment extremity may also have implications for different components of judgment accuracy.[28] To examine this, we calculated accuracy scores for the three studies that were amenable to an analysis of judgment accuracy[29] (Studies 1, 2, and 4). The most commonly used measure of overall accuracy is the quadratic loss function suggested by Brier (1950), which we refer to as the mean probability score ($\overline{PS}$). The procedure for calculating $\overline{PS}$ can be described as follows. Let o_i be an outcome indicator that equals 1 if event o occurs on the ith occasion and 0 otherwise, and let f_i be the forecasted probability of event o on the ith occasion. The mean probability score is given by

$$\overline{PS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{8}$$

where N denotes the total number of trials. Probability scores take a value between 0 and 1, with lower scores indicating greater accuracy.

For Studies 1, 2, and 4, we regressed probability scores onto epistemicness ratings. We conducted all analyses at the trial level using a fractional response model (which accommodates responses that are bounded between 0 and 1; Papke and Wooldridge 1996), with question items treated as fixed effects and standard errors clustered by participants.[30] The first data column of Table 5 provides the average marginal effects for expected probability scores in each study, and Table 6 reports summary statistics for each quartile of rated epistemicness. All three studies return a positive coefficient—suggesting that inaccuracy increased with judged epistemicness—but this relationship was never statistically reliable, as there was considerable noise surrounding these estimates. As indicated in the second data column of Table 5, we also failed to find reliable differences in the proportion of correct judgments (i.e., hit rates) as a function of epistemicness.[31]

These results may at first seem puzzling in light of our robust findings concerning judgment extremity. To summarize, perceptions of epistemicness were associated with more extreme probability judgments but not associated with a reliable increase or decrease in accuracy as measured by Brier scores. As previewed in the introduction, this puzzle is resolved when we partition probability scores into interpretable components.

Table 5. Predicting Different Components of Judgment Accuracy from Perceived Epistemicness

	Probability scores	Proportion correct	Calibration	Resolution
Study 1	0.013	−0.021	0.002***	0.007**
	(0.012)	(0.024)	(0.000)	(0.002)
Study 2	0.005	−0.006	0.002***	0.002**
	(0.005)	(0.010)	(0.000)	(0.001)
Study 4	0.018	0.027	0.003***	0.006*
	(0.013)	(0.032)	(0.001)	(0.002)

Notes. Estimates represent average marginal effects from fractional models, with participant-clustered robust standard errors reported in parentheses. All models include question item fixed effects. For probability and calibration scores, positive coefficients are associated with decreased accuracy. For resolution, positive coefficients are associated with increased accuracy.

***, **, and * indicate significance at 0.001, 0.01, and 0.05, respectively.

Following Murphy (1973), we decompose probability scores as follows:

$$\begin{aligned}\overline{PS} &= \bar{o}(1-\bar{o}) + \frac{1}{N}\sum_{j=1}^{J} n_j(f_j - \bar{o}_j)^2 \\ &\quad - \frac{1}{N}\sum_{j=1}^{J} n_j(\bar{o}_j - \bar{o})^2 \\ &= V + C - R,\end{aligned} \tag{9}$$

in which judgments are grouped into J equivalence classes or bins. In the above equation, n_j is the number of times the judged probability falls into bin j, o_j is the frequency of events in that class, and $\bar{o}$ is the overall relative frequency of the event. For our analyses, judged probabilities were partitioned into bins of 10 (i.e., judgments of 0–0.10, 0.11–0.20, etc.).

The first term on the right-hand side of Equation (9) represents outcome variance (V), or the degree to which the outcome varies from trial to trial. Outcome variance is usually interpreted as an indicator of task difficulty, and therefore it does not directly speak to judgment accuracy. The second term represents judgment calibration (C), or the degree to which actual hit rates deviate from a class of judged probabilities. The third term represents judgment resolution (R), or the degree to which a forecaster reliably discriminates between events that do and do not occur. Whereas calibration provides a measure of how close a judgment is to the truth, resolution provides a measure of the information contained in a forecast. Note that superior performance is represented by lower scores on C and higher scores on R.

Returning to the previous results, we decomposed probability scores to separately analyze calibration and resolution. As before, we conducted analyses at the trial level using a fractional response model with question item fixed effects and standard errors clustered by participants. The average marginal effects

Table 6. Average Probability Score ($\overline{PS}$), Calibration (C), and Resolution (R) Scores as a Function of Judged Epistemicness

	Study 1			Study 2			Study 4		
Epistemicness	$\overline{PS}$	C	R	$\overline{PS}$	C	R	$\overline{PS}$	C	R
Fourth quartile	0.214	0.008	0.070	0.196	0.020	0.059	0.223	0.012	0.042
Third quartile	0.247	0.005	0.056	0.245	0.011	0.047	0.233	0.006	0.027
Second quartile	0.174	0.004	0.053	0.212	0.008	0.036	0.217	0.007	0.029
First quartile	0.192	0.002	0.039	0.228	0.006	0.025	0.203	0.004	0.025

Notes. Epistemic quartiles are ordered from high (fourth quartile) to low (first quartile). For probability and calibration scores, lower numbers indicate greater accuracy. For resolution, higher numbers indicate greater accuracy.

from the regressions are displayed in the last two columns of Table 5, and the results are summarized in Table 6. For all three studies, a consistent pattern emerges. Higher epistemicness ratings were associated with *inferior* performance on calibration but *superior* performance on resolution (because calibration and resolution are scored in opposing directions, positive coefficients imply better calibration but worse resolution).

These results reconcile our finding of no significant association between perceived epistemicness and overall accuracy ($\overline{PS}$) with our finding of a robust association between epistemicness and judgment extremity. On the one hand, heightened perceptions of epistemicness hurt performance by reducing calibration: participants were generally overconfident, and this tendency was exacerbated by more extreme judgments. On the other hand, heightened perceptions of epistemicness helped performance by improving resolution, as participants were more sensitive to differences in evidence strength across events; holding hit rates constant, greater sensitivity should improve discrimination in judgments. Thus, the null effect on overall accuracy reflects the fact that the increase in resolution exhibited by participants who saw events as more epistemic (and less aleatory) was roughly canceled out by a corresponding decrease in calibration. Participants who saw events as primarily epistemic were both more *and* less accurate than participants who saw events as primarily aleatory, depending on the type of accuracy.

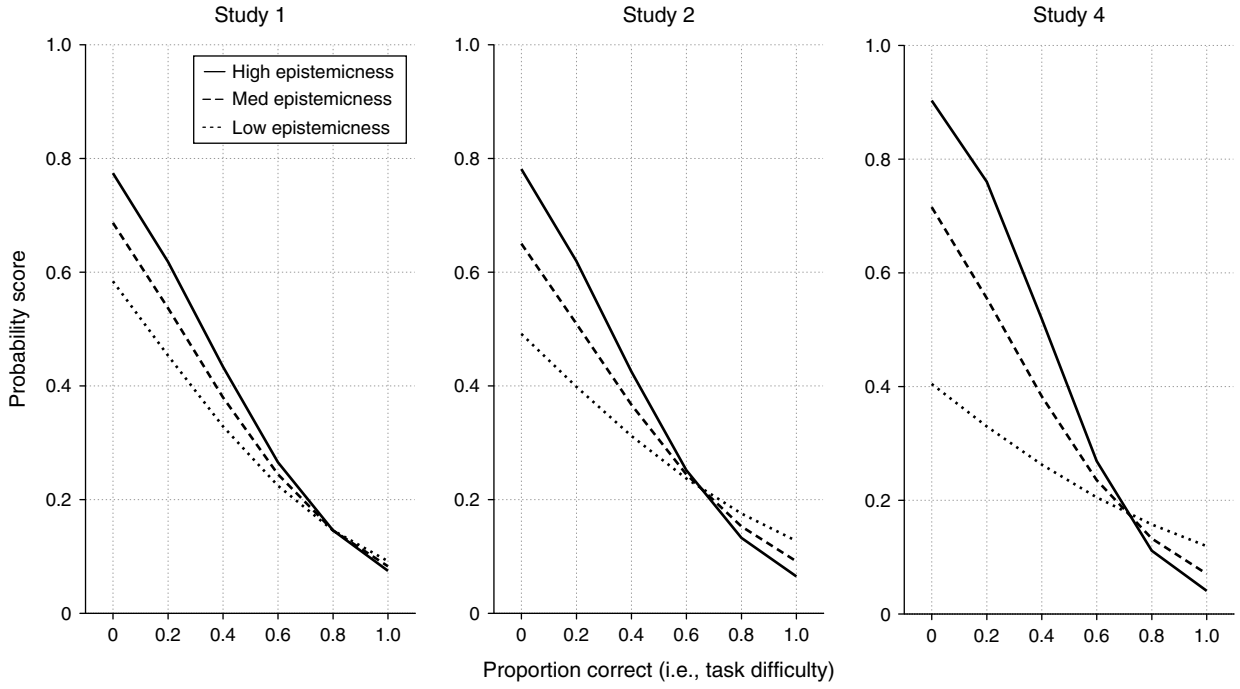
Epistemicness, Overconfidence, and Task Difficulty

If perceptions of epistemicness do not affect hit rates but influence judgment extremity (as shown in the previous section), then we should expect heightened perceptions of epistemic and aleatory uncertainty to improve performance under different task conditions (e.g., easy versus difficult questions; Erev et al. 1994). For task environments that lead to overconfidence—such as difficult questions—the judgment extremity associated with perceptions of high epistemicness should amplify overconfidence (diminish accuracy) whereas the regressiveness associated with perceptions of low epistemicness should attenuate overconfidence (improve accuracy). This pattern should reverse for task environments that typically lead to underconfidence—such as easy questions—where the judgment extremity associated with high epistemicness should reduce underconfidence (improve accuracy), whereas the regressiveness associated with low epistemicness should amplify underconfidence (diminish accuracy). Thus, we would expect overall accuracy ($\overline{PS}$) to be affected by the interaction between perceptions of epistemicness and task difficulty.

To test this prediction, we once again examined the three studies that were amenable to an analysis of judgment accuracy. For each study, we regressed probability scores onto item difficulty (operationalized as the total proportion of correct responses per question), perceptions of epistemicness, and the interaction term. In all analyses, we used a fractional response model with standard errors clustered by participants. The results are depicted in Figure 4, where predicted mean probability scores are plotted against task difficulty at low, medium, and high levels of perceived epistemicness (one standard deviation below the mean, at the mean, and one standard deviation above the mean, respectively). The graphs show a general downward trend in expected probability scores as the proportion of correct responses increases, reflecting the fact that accuracy improves as task questions become less difficult. More important, in all three cases we found a reliable interaction effect that conforms to the expected pattern of results ($p = 0.029$ for Study 1 and $p < 0.001$ for Studies 2 and 4). Perceptions of greater epistemicness were usually associated with superior calibration for easy questions (lower $\overline{PS}$) but inferior calibration for difficult questions (higher $\overline{PS}$). As predicted, differences in perceptions of epistemic and aleatory uncertainty resulted in enhanced accuracy under different task conditions.

An interesting avenue for future research will be to determine whether insights gleaned from the epistemic–aleatory distinction can be leveraged to formulate interventions or elicitation techniques to improve judgment accuracy. For example, the current results suggest that for domains in which forecasters typically display overconfidence, one may wish

Figure 4. Accuracy as a Function of Task Difficulty and Judged Epistemicness



to highlight the aleatory uncertainty inherent to the judgment task, whereas for domains in which forecasters typically display underconfidence, one may wish to highlight the epistemic uncertainty inherent to the judgment task. We note that an established technique for reducing overconfidence has been to prompt disconfirmatory thinking ("consider the opposite")—when individuals are first asked to think of how an event could have turned out differently than expected, their subsequent judgments tend to be less overconfident (Arkes et al. 1988, Koriat et al. 1980, Hoch 1985). We suspect that considering alternative outcomes increases the salience of aleatory uncertainty—it makes the target event appear more random and less predictable—which in turn leads to more regressive judgments and therefore attenuates overconfidence. Although existing research has not to our knowledge examined interventions for reducing systematic underconfidence, we expect procedures that highlight the inherent knowability of an uncertain event (i.e., increasing the salience of epistemic uncertainty) may be a fruitful approach.

Variability in Assessments of Evidence Strength

Brenner and colleagues (Brenner 2003, Brenner et al. 2005) developed a random support model of subjective probability that provides an alternative approach to modeling variability in judgment extremity. Random support theory posits that judgment extremity arises from variability in the evidence that a judge recruits for the same hypothesis on different occasions. The idea is that support is randomly drawn from a log-normal distribution, with greater variability in this distribution resulting in more extreme judgment. Brenner (2003) provided empirical evidence for this interpretation by showing that variability in support distributions (measured using strength ratings as we have done) were strongly associated with more extreme probability judgments. This finding motivated us to reexamine our data to see whether between-subject variability in strength ratings (which following Brenner 2003, we used as an empirical proxy for within-subject variance in support distributions) could account for our results.

Study 4 allows for the most direct test of the random support model, as this study was conducted between participants and held the strength elicitation format constant across experimental task conditions. We conducted robust tests of variance with adjustments made for clustered data (Levene 1960, Iachine et al. 2010). For completeness, we conducted tests using the mean absolute difference, median absolute difference, and 10% trimmed mean absolute difference in strength ratings, and we performed these tests on the variance in strength ratios, $\hat{s}(A)/\hat{s}(B)$, as well as separately for variance in focal and alternative strength ratings ($\hat{s}(A)$ and $\hat{s}(B)$, respectively). For all tests, we failed to find any reliable differences across conditions: p-values were always above 0.10 (with an average p-value of 0.46 across all tests), and the observed R^2 from every test was always less than 0.01. In short, our experimental conditions had a reliable influence on judgment extremity in a way that could not be accounted for by differences in the variability of strength ratings.

Knowledge and Sensitivity to Evidence Strength

Our analysis of the relationship of raw strength ratings and judged probabilities relies on the original formulation of support theory. However, support theory does not directly account for the fact that people vary in their levels of knowledge or expertise. For example, people give more regressive probability estimates when they feel relatively ignorant about the task at hand (e.g., Yates 1982) and often report probabilities of 1/2 when they feel completely ignorant (Fischhoff and Bruine De Bruin 1999). It may be that levels of subjective knowledge interact with the effects we report here. For example, if participants feel ignorant or uninformed about a task, they are likely to provide highly regressive judgments regardless of the degree of perceived epistemicness. More generally, one might suppose that the impact of perceived epistemicness on judgment extremity is attenuated in situations where people feel relatively ignorant and amplified in situations where they feel relatively knowledgeable (Fox and Ülkümen 2011).

Future work can explore this prediction by using an extension of support theory that incorporates reliance on *ignorance prior probabilities* (i.e., probabilities that assign equal credence to every hypothesis into which the state space is partitioned; Fox and Rottenstreich 2003, Fox and Clemen 2005, See et al. 2006). For instance, Fox and Rottenstreich (2003) propose a model in which probability judgments are represented as a convex combination of evidence strength and the ignorance prior probability (i.e., $1/n$ for n-alternative questions). In this model the judged odds $R(A,B)$ that hypothesis A obtains rather than its complement B are given by

$$R(A,B) = \left[\frac{n_A}{n_B}\right]^{1-\lambda}\left[\frac{\hat{s}(A)}{\hat{s}(B)}\right]^{k'\lambda}. \tag{10}$$

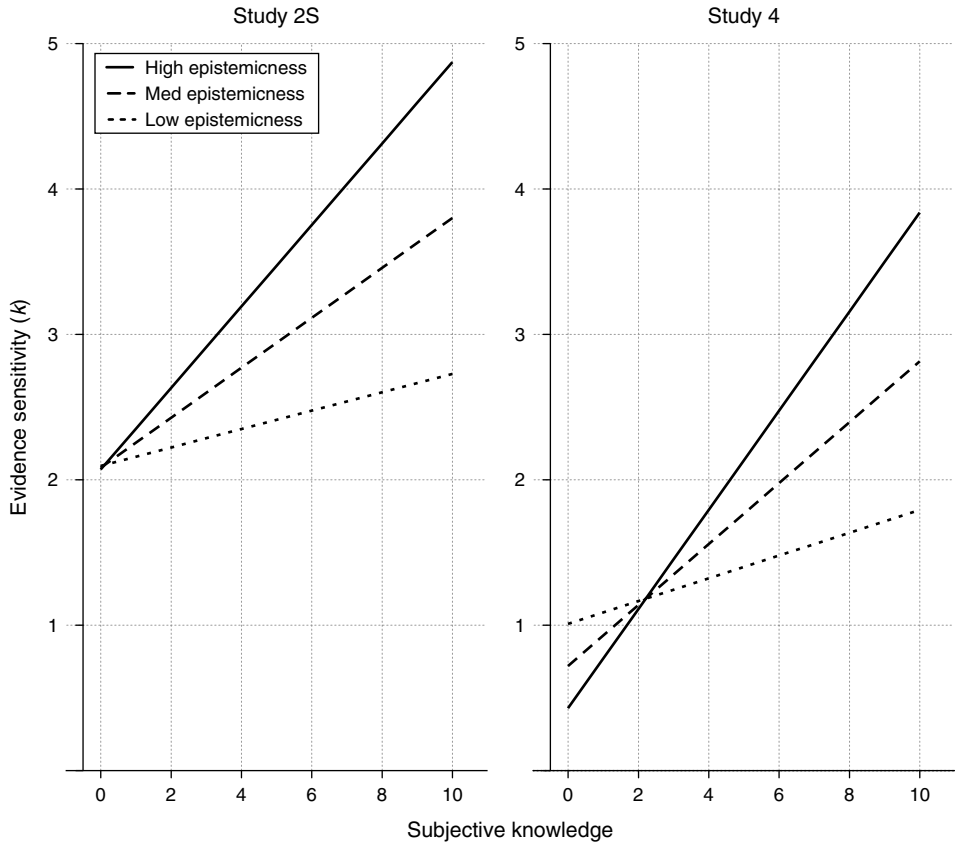
The second expression on the right-hand side of Equation (10) represents the balance of support as measured by raw strength ratings, akin to the original support theory formulation presented in Equation (3). The first expression on the right-hand side represents the ignorance prior (in odds format) for the focal hypothesis A relative to the alternative hypothesis B. For two-alternative questions this implies odds of 1:1, for three-alternative questions this implies odds of 1:2, and so forth. The parameter λ represents the proportion of weight afforded the ignorance prior relative to the support ratio, and takes a value between 0 and 1. As λ approaches 1, more weight is placed on the balance of evidence (i.e., support values); as λ approaches 0, judgments converge toward the ignorance prior. One can interpret λ as an indicator of subjective knowledge. When people feel relatively ignorant, they are likely to afford more weight on the ignorance prior; when people feel relatively knowledgeable, they tend to give less weight to the ignorance prior and increasingly rely on subjective impressions of relative evidence strength. Finally, k' measures (partition-independent) sensitivity to differences in evidence strength (note that k in Equation (3) has now been decomposed into λ and k').

The ignorance prior model makes a clear prediction concerning the interaction between subjective knowledge and perceptions of epistemicness on sensitivity to evidence: the tendency for evidence sensitivity to increase with perceived epistemicness should be amplified when participants are more knowledgeable (i.e., when they rely less on the ignorance prior) and should be attenuated when participants are less knowledgeable (i.e., when they rely more on the ignorance prior).

For exploratory purposes, we asked participants at the end of our studies to rate their level of knowledge[32] for each judgment domain. For the two studies in which we asked participants to rate their knowledge separately for each domain or task and where we could statistically estimate evidence sensitivity over participants for those domains/tasks (Studies 2S and 4), we examined the interaction between epistemicness and subjective knowledge on evidence sensitivity. For each study, we recovered sensitivity coefficients for each participant and then regressed these estimates onto each participants' epistemicness ratings, self-reported knowledge, and the interaction term. In Figure 5, we plot for each study evidence sensitivity for low, medium, and high epistemicness ratings (one standard deviation below the mean, at the mean, and one standard deviation above the mean) across the range of subjective knowledge ratings. As predicted by the ignorance prior model (and anticipated by Fox and Ülkümen 2011), we see a general "fanning-out" effect as knowledge increases—differences in evidence sensitivity between high and low perceived epistemicness were most pronounced when knowledge was high. The interaction term between rated epistemicness and task knowledge was in the predicted direction for both studies (p-values were 0.104 and 0.055 for Studies 2S and 4, respectively), and the overall p-value was 0.035 by Fisher's combined probability test. Additional evidence consistent with the notion that the level of subjective knowledge moderates the relationship between epistemicness and evidence sensitivity, based on internal analyses of Studies 1 and 2, are reported in the supplementary materials.

Although consistent with the ignorance prior model, these results should be treated as tentative. The measurement approach for subjective knowledge was considerably more coarse (i.e., a single-item self-report measure) than were values for sensitivity to evidence strength (which were derived from multiple trials of judgments and strength ratings). Future work could more rigorously test the knowledge amplification prediction by independently manipulating the ignorance

Figure 5. Sensitivity to Evidence Strength as a Function of Subjective Knowledge and Judged Epistemicness



prior alongside the measurement of probability judgments and strength ratings (for an example of this approach that did not include epistemicness ratings, see See et al. 2006).

Interpretations of Epistemic Extremity

In this paper, we have suggested that the tendency for more extreme judgments under more epistemic uncertainty is driven by a tendency to see the balance of evidence as more diagnostic of outcomes under these conditions. We note that perceptions of epistemic and aleatory uncertainty are subjective and may not necessarily agree with the actual predictability or randomness of events in the task environment. For one thing, individuals sometimes fail to appreciate the stochastic nature of events they perceive as highly epistemic. Prior research has noted that overconfidence often occurs because people formulate judgments that are conditional on their beliefs or model of the world being true, and they fail to acknowledge the possibility that their interpretations and knowledge may be off (Dunning et al. 1990, Griffin et al. 1990, Trope 1978). When uncertainty is perceived to be more aleatory, an individual may be more likely to make "inferential allowances" (i.e., more regressive judgments) because viewing events as aleatory highlights the role of chance processes in determining an outcome. Under epistemic uncertainty, however, that same individual may instead focus on what she knows and thus fail to appreciate that her beliefs could be incorrect or partly determined by stochastic factors (e.g., that the information retrieval process underlying impressions of evidence strength is subject to random noise). In short, for epistemic but not aleatory uncertainty, people may confuse absence of doubt in their beliefs about an event with the belief that an event is undoubtedly true.

Another possible and complementary mechanism driving increased judgment extremity under more epistemic uncertainty is the notion that purely epistemic events (e.g., the correct answer to a trivia question) are either true or false whereas pure aleatory events (e.g., whether a fair die will land on a prime number) have intermediate propensities. To illustrate, consider the purely epistemic question of whether one country is geographically larger than another. Given a person's impression of the relative sizes of these two countries, his judged probability of this event should quickly approach 0 or 1 as this impression becomes increasingly distinct. Next, consider a purely aleatory event such as whether a roulette wheel will land on one of the numbers that another person has bet on. This question entails an event that has a "true" propensity that may lie anywhere along the $[0, 1]$ probability

interval. In this case, even if the other person has bet on nearly every available number, her judged probability should remain less than 1. We conjecture that because of the principles of stimulus-response compatibility (Fitts and Seeger 1953, Wickens 1992), events that are seen as more epistemic may more naturally tend toward 0 or 1 than events that are seen as more aleatory.

Support for this notion can be found in a previous finding from Ronis and Yates (1987), in which participants expressed judgments of complete certainty (judged probabilities of 0 or 1) on 25% of their responses to trivia questions, compared with only 1.3% of responses to basketball games. Extending this finding, our own data show a consistent pattern that 0 and 1 responses are more common for events that were rated as more epistemic (see the final column of Table 2).

Conclusion

Experts and laypeople confront uncertainty every day. Whether evaluating an investment, forecasting a geopolitical outcome, or merely assessing whether it is safe to cross the street, individuals must evaluate the likelihood of events that can be construed to varying degrees as knowable or random. In this paper, we have documented a general tendency for judgments to be more extreme when they are seen as more epistemic and less extreme when they are viewed as more aleatory. We have observed that such differences in judgment extremity may also help to explain a number of stylized findings from the literature on judgment accuracy and overconfidence, and consequently, they may inform procedures and elicitation techniques for improving judgment accuracy.

Acknowledgments

The authors thank Dan Walters and Carsten Erner for useful comments and suggestions on a previous version of this manuscript. The authors also thank Adam Waytz for his assistance with data collection on their initial NBA demonstration study.

Endnotes

[1] Another way to think about the distributional–aleatory link is that an aleatory mind-set implies distributional thinking: it would be difficult to consider chance variability without thinking about a range of possible instances that are similar in some respects. Conversely, distributional thinking may promote an aleatory mind-set: when one considers the target event as an instance of a distribution of similar events, this may highlight the possibility of random variability among outcomes.

[2] We recruited fans from online basketball forums and through social networking websites (99% male, mean age = 31 years, range: 21–50 years). In return for their participation, we entered all participants into a raffle to receive an NBA jersey of their choice. Participants reported watching a median of 3 NBA basketball games per week, 45 games for the season, and 10 hours per week listening to or watching sports commentary about NBA basketball. The order of games was presented in a randomized order for each participant, as was the team designated as focal for each game (e.g., whether participants judged the probability that the Bulls or the Pistons would win).

[3] Since we make ex ante directional predictions, we report p-values from one-tailed tests throughout the paper unless otherwise noted.

[4] We also collected data on perceived strength of the teams in question (which was not affected by the framing manipulation, as expected) and the perceived epistemicness/aleatoriness of a typical professional basketball game (which did not register a significant difference by condition, perhaps because the question was framed too generically or because we relied on an abbreviated scale as a result of time constraints).

[5] This included all first-round games, excluding the four play-in games that had yet to be played when we ran the study. Apart from the four play-in games (two 16th seed and two 11th seed teams), this left four each of every strength matchup (1 versus 16, 2 versus 15, 3 versus 14, etc.).

[6] This format for eliciting judged probabilities, sometimes referred to as the designated form (Liberman and Tversky 1993), can be contrasted with a forced-choice format that prompts participants to first choose a team and then provide a probability judgment from 0.5 to 1. We chose the designated form for eliciting beliefs because it allows us to distinguish over-*extremity* (the tendency to provide judgments that are too close to 0 or 1) from over-*prediction* (the tendency to overestimate the likelihood of all events). Formats such as two-alternative forced-choice questions cannot distinguish between the two (see Brenner et al. 2005).

[7] In prior development of the EARS (Fox et al. 2016), as well as the current application, the scale loads onto two separate dimensions. However, in the present context—in which we predict complementary effects of greater judgment extremity under increasing epistemic uncertainty and decreasing aleatory uncertainty—we treat them as a single dimension (by reverse-coding aleatory scale items) for simplicity and ease of exposition. We obtain qualitatively identical results if we separately analyze epistemic and aleatory subscales in Studies 1–3; in Study 4 we supplement our analysis of the unitary EARS with an analysis of the subscales (see Endnote 27).

[8] To give a sense of the magnitude of this effect, the judged probability that a given team would win their matchup increased by an average of 4.2 percentage points per increase in seed ranking.

[9] The average marginal effect represents the instantaneous rate of change in the dependent variable as a function of a predictor variable, averaged over all observations. In the result reported above, for example, if the rate of change was constant, then we would expect a 5.5-percentage-point increase in complete certainty responses for every one-unit increase in epistemicness ratings.

[10] Another reason to examine evidence sensitivity is that, within a single domain, parity between hypotheses might otherwise affect perceptions of epistemicness and aleatoriness. For instance, if two teams are rated as equally strong, then a judge might view the outcome of the game as more aleatory (random). Conversely, if two teams are extremely unbalanced, the outcome may be seen as more epistemic (knowable). Our measure of evidence sensitivity explicitly controls for differences in parity of evidence strength, thereby removing this potential confound.

[11] The emphasis on hypotheses, rather than events, allows for the possibility that different descriptions of the same event can elicit different probabilities (i.e., the framework is nonextensional). In the present paper we assume a canonical description of events, so this distinction will not be relevant.

[12] Because hypothetical support, $s(\cdot)$, is a ratio scale, any strength measure used as a proxy, $\hat{s}(\cdot)$, ought to be evaluated on a ratio scale

as well. Thus, following Tversky and Koehler (1994), our strength elicitation protocol explicitly instructs participants to assign strength in proportion to the strongest event so that, for example, a team that is strongest in a set would be assigned a strength of 100 and a team that is seen as half as strong would be assigned a strength of 50. Although not all strength measures have a natural 0 point to establish a ratio scale (e.g., what does it mean for a team to have zero strength?), past studies using support proxies such as team strength have established that participants can adequately scale strength so that the model fits quite well (e.g., Tversky and Koehler 1994, Koehler 1996). Moreover, Fox (1999) explicitly tested the monotonicity conditions using ratings of team strength and found that they both held quite well.

[13] A related observation about the interpretation of k was made by Koehler (1996, p. 20): "One speculation is that the value of k may reflect the relative predictability of the outcome variable in question. Thus, for example, considerably lower values of k would be expected if subjects were asked to judge the probability that the home team will score first in the game (rather than that the home team will win the game) because this variable is generally less predictable." Here, we suggest that k tracks beliefs about the nature of the uncertainty—the extent to which the outcome is epistemic/aleatory—rather than mere predictability.

[14] For all studies conducted on MTurk, we restricted the sample to U.S. participants. Furthermore, we used software that excluded participants who had completed any of the previous experiments (Goldin and Darlow 2013).

[15] Although participants were allowed to express very small strength numbers, in no case did we present them with hypotheses that could plausibly be associated with vacuous strength.

[16] In Study 2 there is only one question per domain, so including a question fixed effect term would be redundant with the treatment term.

[17] Because the regression analysis required transforming probability judgments into log odds, responses of 0 and 1 were compressed by substituting them with $0.5/N$ and $[N - 0.5]/N$, respectively, for sample size N, as suggested by Smithson and Verkuilen (2006).

[18] Although not the primary focus of the current research, the intercept in our model can also be interpreted as an index of response bias. A key assumption of support theory is *binary complementarity*, which states that judgments to two-alternative outcomes are additive (i.e., $p(A, B) + p(B, A) = 1$). Thus, if binary complementarity holds, then we should not expect any appreciable degree of response bias (i.e., the intercept should not differ substantially from 0). Consistent with support theory, we found that the intercept term in all studies, with the exception of Study 3, did not reliably differ from 0. More direct tests of binary complementarity are provided in the supplementary materials.

[19] An estimated measure of evidence sensitivity for each participant per domain was not possible, since participants were merely asked to provide a single judgment per domain. Note that the analytic measure requires us to exclude a small number of observations (n = 20) where participants judged the focal and alternative targets to be equally strong (since this places a value of 0 in the denominator).

[20] Or they were more reliably evaluated on a ratio scale—see Endnote 11.

[21] Note that the design of Study 3 bears an interesting relationship to our short demonstration (discussed in the introduction) in which participants judged probabilities that various teams would win specific NBA matchups. In both that study and in Study 3 we predict judgments should be especially regressive when an event is viewed as an instance from a distribution of similar events. Highlighting the distributional nature of an event should focus attention on the possible variation across similar occasions and lead individuals to view an event as more aleatory. In our NBA demonstration, this was accomplished by prompting individuals to think about a particular game as one of a series of contests between the two teams. For Study 3, by contrast, we prompt aleatory thinking by asking individuals to consider city temperatures for an "arbitrarily selected day" from the previous year. We surmise that the arbitrarily selected single-day task is more apt to be construed as an instance from a larger set of similar events than judgments about average city temperatures over a year.

[22] The reason we used the language "arbitrary day" rather than "randomly selected day" is that the latter might cause a demand effect when participants rate events on the EARS (one of the items uses the word "random").

[23] These results also hold when analyzing evidence sensitivity over participants or over items, rather than over trials as we did in the foregoing analysis.

[24] Epistemicness ratings were measured at the trial level for each participant. Similar to our previous analyses, we include question fixed effects and participant random effects in the model.

[25] The ordering of the two binary prediction blocks also did not systematically affect judgment extremity or evidence sensitivity.

[26] Similar to Study 3, these results also hold when analyzing evidence sensitivity over participants and over items, rather than over trials as we did in the foregoing analysis.

[27] We separately examined epistemic and aleatory items from the scale and found that the correlation between these two indices was weaker than in all of our other studies ($r = -0.21$). We proceeded to analyze each subscale separately and found no reliable difference in ratings on the epistemic subscale (means were 4.72 and 4.81, respectively; $p = 0.69$) but a significant difference in the expected direction on the aleatory subscale (means were 4.10 and 4.55, respectively; $p = 0.028$). That is, participants viewed questions as higher in aleatory uncertainty for the random prediction prime than for the pattern detection prime. Furthermore, we examined the relationship between epistemicness and sensitivity to evidence strength using only the aleatory subscale, and we observed an even stronger effect than that reported above using the full EARS ($p = 0.005$ for the interaction term).

[28] Because when initially undertaking this analysis we did not make ex ante directional predictions about judgment accuracy, we report two-tailed test statistics for these analyses.

[29] Study 3 was excluded from the analysis because half of the questions involved estimating upcoming temperatures for arbitrarily selected days from the previous year, which poses difficulties for calculating accuracy scores. The most natural way to code outcomes for this task would be to use the base rate over the estimation interval (e.g., the proportion of warmer days in city A over city B during a one-year period), but doing so dramatically reduces the outcome variance compared with yearlong average questions where outcomes are coded as 0 or 1. Thus, interpreting any differences in judgment accuracy across domains is problematic because perceptions of epistemicness will be conflated with task difficulty (i.e., outcome variability). Study 2S was not included in the analysis for the same reason and also because judgments of basketball games—which made up one-third of the stimulus items in the study—were based on possible matchups that were not always realized (e.g., "Suppose the San Antonio Spurs play the Philadelphia 76ers in the NBA finals…") and thus could not be scored.

[30] When analyzing judgment extremity earlier on, we chose to report estimates from linear models rather than from fractional response models because both approaches provided similar results and the former was simpler to convey. We chose to use fractional response models here because, unlike our earlier analyses, many of the observations for components of judgment accuracy (such as calibration and resolution scores) lie at or near the scale boundary of 0. As a

result, using a linear model produced estimates that were often out of range. We note that using linear models instead returns similar results to those reported above.

[31] Hit rates were calculated for each judgment by assigning a score of 1 when participants provided a judged probability above 0.50 and the event occurred, or a probability below 0.50 and the event failed to occur. Participants were assigned a score of 0 when judged probability was below 0.50 and the event occurred, or above 0.50 and the event failed to occur. For responses of 0.50, we randomly assigned responses as correct or incorrect (see Ronis and Yates 1987).

[32] For Study 4, knowledge was assessed on a 11-point scale from 0 (not knowledgeable at all) to 10 (very knowledgeable). For Study 2S, we assessed knowledge in a similar manner but using a 100-point scale, which we subsequently transformed (by multiplying responses by 0.1) for purposes of comparison.

References

Arkes HR, Faust D, Guilmette TJ, Hart K (1988) Eliminating the hindsight bias. *J. Appl. Psych.* 73(2):305–307.

Brenner LA (2003) A random support model of the calibration of subjective probabilities. *Organ. Behav. Human Decision Processes* 90(1):87–110.

Brenner L, Griffin D, Koehler DJ (2005) Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organ. Behav. Human Decision Processes* 97(1):64–81.

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

Carlson BW (1993) The accuracy of future forecasts and past judgments. *Organ. Behav. Human Decision Processes* 54(2):245–276.

Dunning D, Griffin DW, Milojkovic JD, Ross L (1990) The overconfidence effect in social prediction. *J. Personality Soc. Psych.* 58(4):568–581.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–-527.

Fischhoff B, Beyth R (1975) I knew it would happen: Remembered probabilities of once-future things. *Organ. Behav. Human Performance* 13(1):1–16.

Fischhoff B, Bruine De Bruin W (1999) Fifty–fifty = 50%? *J. Behav. Decision Making* 12(2):149–163.

Fitts PM, Seeger CM (1953) S-R compatibility: Spatial characteristics of stimulus and response codes. *J. Experiment. Psych.* 46(3): 199–210.

Fox CR (1999) Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psych.* 38(1):167–189.

Fox CR, Clemen RT (2005) Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Sci.* 51(9):1417–1432.

Fox CR, Rottenstreich Y (2003) Partition priming in judgment under uncertainty. *Psych. Sci.* 14(3):195–200.

Fox CR, Tversky A (1998) A belief-based account of decision under uncertainty. *Management Sci.* 44(7):879–895.

Fox CR, Ülkümen G (2011) Distinguishing two dimensions of uncertainty. Brun W, Keren G, Kirkeben G, Montgomery H, eds. *Perspectives on Thinking, Judging, and Decision Making: A Tribute to Karl Halvor Teigen* (Universitetsforlaget, Oslo, Norway), 21–35.

Fox CR, Tannenbaum D, Ülkümen G (2016) The empirical case for distinguishing two dimensions of subjective uncertainty. Working paper, University of California, Los Angeles, Los Angeles.

Gaissmaier W, Schooler LJ (2008) The smart potential behind probability matching. *Cognition* 109(3):416–422.

Goldin G, Darlow A (2013) TurkGate (version 0.4.0). [Software.] Retreived December 12, 2012, http://gideongoldin.github.io/TurkGate/.

Goodnow JJ (1955) Determinants of choice-distribution in two-choice situations. *Amer. J. Psych.* 68(1):106–116.

Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Griffin DW, Dunning D, Ross L (1990) The role of construal processes in overconfident predictions about the self and others. *J. Personality Soc. Psych.* 59(6):1128–1139.

Hacking I (1975) *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference* (Cambridge University Press, Cambridge, UK).

Herrnstein RJ (1997) *The Matching Law: Papers in Psychology and Economics* (Harvard University Press, Cambridge, MA).

Hoch SJ (1985) Counterfactual reasoning and accuracy in predicting personal events. *J. Experiment. Psych.: Learn., Memory, Cognition* 11(4):719–731.

Howell WC, Kerkar SP (1982) A test of task influences in uncertainty measurement. *Organ. Behav. Human Performance* 30(3):365–390.

Iachine I, Petersen HC, Kyvik KO (2010) Robust tests for the equality of variances for clustered data. *J. Statist. Comput. Simulation* 80(4):365–377.

Keren G (1991) Calibration and probability judgements: Conceptual and methodological issues. *Acta Psych.* 77(3):217–273.

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

Koehler DJ (1996) A strength model of probability judgments for tournaments. *Organ. Behav. Human Decision Processes* 66(1): 16–21.

Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for confidence. *J. Experiment. Psych.: Human Learn. Memory* 6(2):107–18.

Levene H (1960) Robust tests for equality of variances. Olkin I, ed. *Contributions to Probability and Statistics*, Vol. 2 (Stanford University Press, Palo Alto, CA), 278–292.

Liberman V, Tversky A (1993) On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psych. Bull.* 114(1):162–173.

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

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

Murphy AH (1973) A new vector partition of the probability score. *J. Appl. Meteorol.* 12(4):595–600.

Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J. Appl. Econometrics* 11(6):619–632.

Robinson EJ, Rowley MG, Beck SR, Carroll DJ, Apperly IA (2006) Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development* 77(6):1642–1655.

Ronis DL, Yates JF (1987) Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organ. Behav. Human Decision Processes* 40(2):193–218.

Rottenstreich Y, Tversky A (1997) Unpacking, repacking, and anchoring: Advances in support theory. *Psych. Rev.* 104(2):406–415.

Savage LJ (1954) *The Foundations of Statistics* (John Wiley & Sons, New York).

See KE, Fox CR, Rottenstreich YS (2006) Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *J. Experiment. Psych.: Learn., Memory, Cognition* 32(6): 1385–1402.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych. Sci.* 22(11):1359–1366.

Smithson M, Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psych. Methods* 11(1):54–71.

Tenney ER, MacCoun RJ, Spellman BA, Hastie R (2007) Calibration trumps confidence as a basis for witness credibility. *Psych. Sci.* 18(1):46–50.

Trope Y (1978) Inferences of personal characteristics on the basis of information retrieved from one's memory. *J. Personality Soc. Psych.* 36(2):93–105.

Tversky A, Koehler DJ (1994) Support theory: A nonextensional representation of subjective probability. *Psych. Rev.* 101(4):547–567.

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

Unturbe J, Corominas J (2007) Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology* 21(5):621–630.

Volz KG, Schubotz RI, von Cramon DY (2004) Why am I unsure? Internal and external attributions of uncertainty dissociated by fMRI. *Neuroimage* 21(3):848–857.

Volz KG, Schubotz RI, von Cramon DY (2005) Variants of uncertainty in decision-making and their neural correlates. *Brain Res. Bull.* 67(5):403–412.

von Mises R (1957) *Probability, Statistics, and Truth* (Dover Publications, New York).

Wickens CD (1992) *Engineering Psychology and Human Performance*, 2nd ed. (HarperCollins, New York).

Wolford G, Newman SE, Miller MB, Wig GS (2004) Searching for patterns in random sequences. *Canadian J. Experiment. Psych.* 58(4):221–228.

Wright G (1982) Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psych.* 52(1): 165–174.

Wright G, Ayton P (1987) Task influences on judgemental forecasting. *Scand. J. Psych.* 28(2):115–127.

Wright G, Wisudha A (1982) Distribution of probability assessments for almanac and future event questions. *Scand. J. Psych.* 23(1): 219–224.

Yates JF (1982) External correspondence: Decompositions of the mean probability score. *Organ. Behav. Human Performance* 30(1): 132–156.