Judgment Extremity and Accuracy under Epistemic versus Aleatory Uncertainty

David Tannenbaum and Craig R. Fox
Anderson School of Management, UCLA

Gülden Ülkümen
Marshall School of Business, USC

Abstract

We show that people make more extreme probability judgments (i.e., closer to 0 or 1) for events they see as entailing more epistemic (knowable) uncertainty and less aleatory (random) uncertainty. We demonstrate this pattern when there is consensus concerning the balance of evidence favoring the target event (Study 1), across a range of domains in which we control for rated evidence strength (Studies 2A and 2B), when events differ only in the degree of randomness with which they are selected (Study 3), and when we prime participants to see events as more epistemic or more aleatory (Study 4). Decomposition of Brier scores suggests that the greater extremity of more epistemic events reflects a trade-off between higher resolution and lower 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.

Introduction

Judgment under uncertainty entails two challenges — what to believe, and how strongly to hold those beliefs. Determining an appropriate strength of beliefs is critical for a wide range of decisions by both laypeople and experts. For instance, jurors in U.S. criminal cases must not only decide whether they think a suspect is guilty, but also whether they are guilty beyond “reasonable doubt.” Physicians are frequently called on to advise their patients not only what course of treatment to pursue, but also how likely that treatment is to succeed. Vacationers confronting a decision to purchase travel insurance must determine not only whether a trip cancellation is possible, but also the likelihood that a cancellation will occur. 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*, or the degree to which probabilistic beliefs approach 0 or 1. A well-established literature finds that people are prone to excessive overconfidence 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 of events relative to their empirical frequencies defines both overconfidence (judgments that are too extreme) and underconfidence (judgments that are not sufficiently extreme). Furthermore, judgment extremity is a critical driver of willingness to act under uncertainty (e.g., Fox and Tversky, 1998). Thus, understanding the psychological processes that give rise to judgment extremity can shed light on both judgment accuracy and decisions under uncertainty.

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

Across five studies, we find that judgments are more extreme for events that are perceived to be more epistemic (knowable) and more regressive for events that are perceived to be more aleatory (random). Prior research suggests that judged probabilities are well predicted by the balance of assessed evidence for versus against a hypothesis, which is mapped onto degrees of belief (Tversky and Koehler, 1994; Rottenstreich and Tversky, 1997; Fox, 1999). We find that observed differences in judgment extremity under epistemic versus aleatory uncertainty reflect differences in the mapping of evidence strength onto judged probabilities, and do not appear to affect the perceived evidence for the hypotheses under consideration.

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 hypotheses, 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 framework to an analysis of judgment accuracy. We demonstrate that perceptions of epistemic and aleatory uncertainty enhance and diminish different components of judgment accuracy, and also have implications for improving accuracy in task environments that lead to systematic overconfidence versus underconfidence. We conclude with a number of additional theoretical observations.

Epistemic versus Aleatory Judgment Under Uncertainty

Most theories of judgment and decision-making construe uncertainty as a unitary construct. For instance, in Bayesian decision theories (e.g., Savage, 1954) subjective probabilities are treated as degrees of belief, regardless of their source. Meanwhile, frequentistic accounts of probability (e.g., Von Mises, 1957) restrict their attention to situations in which there are stable long-run relative frequencies of classes of events. Fox and Ülkümen (2011) reviewed evidence 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 or propositions perceived to be inherently knowable (epistemic uncertainty) from those perceived to be inherently random or stochastic (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 operationalize 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, 4-6 year old children tend to behave differently when facing chance events yet to occur (in which aleatory or random uncertainty is presumably salient) versus chance events that have already been resolved but not yet revealed (in which epistemic or knowable uncertainty is presumably salient; Robinson et al., 2006). Brain imaging studies (Volz et al., 2005, 2004) found distinct activation patterns when participants learned about events whose outcomes were determined in a stochastic (presumably aleatory-salient) manner compared to events determined in a rule-based (presumably epistemic-salient) manner. More recently, Ülkümen et al. (2014) found that people tend to associate words like "chance" and "likelihood" (e.g., "I think there's a 80% chance" or "I believe that there is a high likelihood") with events that are characterized by aleatory uncertainty whereas they tend to associate words like "sure" and "confident" (e.g., "I am 80% sure" or "I am highly confident") with events that are characterized by epistemic uncertainty.

It stands to reason that perceptions of epistemicness and aleatoriness should also affect judgment extremity. First, consider purely epistemic (knowable) uncertainty such as the question of whether one country is geographically larger than another country. This question entails an event that is either true or false. Thus, 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. For example, given that a typical person has a distinct impression that Spain is smaller than Russia but larger than Portugal, his judged probabilities that Spain is larger than Russia and Portugal should converge to 0 and 1, respectively. Next, consider a purely aleatory (random) event such as whether a roulette wheel will land on one of the numbers on which a person has bet. 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 she has placed bets on the large majority of available numbers, her judged probability should remain less than one.

More interesting is the case of events that entail a mixture of both epistemic and aleatory uncertainty. In this case, holding perceived evidence strength constant, the more a person sees the event as epistemic (knowable) the more he or she may tend toward 0 or 1, and the more a person sees the event as aleatory (random) the more he or she will regress toward intermediate probabilities. To illustrate, suppose I am predicting the winner of a chess match and the winner of a game of darts, and assume that I view chess as more predictable and less random than darts. Perhaps this is partly due to the fact that I recognize that darts entails more inherent variability in performance so that impressions of relative strength provide a less diagnostic cue. Thus, I will generally assign higher probabilities to the stronger player winning in chess than darts, and as the perceived strength imbalance increases I will tend toward more extreme probabilities faster for chess than darts. A similar logic also applies to the same event if multiple decision makers are considered. For example, two people could differ in the degree of epistemic and aleatory uncertainty they impute to the outcome of a game of darts, and we would expect a similar difference to emerge in the extremity of their probability judgments.

Existing research on judgment overconfidence accords with the foregoing hypothesis that judgments tend to be more extreme for events that are more epistemic (knowable) and less aleatory (random). As other researchers have observed (e.g., Wright and Ayton, 1987), studies examining judgment accuracy typically use general knowledge items such as trivia questions (see Lichtenstein et al., 1982). While general knowledge questions have the advantage of being immediately verifiable — researchers do not have to wait for the outcomes to occur to score responses — it has also been found that overconfidence is often highest for these types of judgment tasks. Note that uncertainty associated with responding to trivia questions will tend to be experienced as purely epistemic, and therefore we hypothesize this is where judgment extremity will be most pronounced. When the accuracy of judgment is examined in other domains that appear to entail more aleatory uncertainty, such as future geopolitical events or sporting matches, overconfidence appears to be considerably attenuated (Carlson, 1993; Fischhoff and Beyth, 1975; Howell and Kerkar, 1982; Ronis and Yates, 1987; Wright, 1982; Wright and Wisudha, 1982). For example, Ronis and Yates (1987) compared probability judgments of trivia questions to upcoming professional basketball games, and Wright and Wisudha (1982) compared probability judgments of trivia questions to then-future events. In both cases, participants provided relatively less extreme judgments, and displayed less overconfidence, when judging the outcomes of basketball games or future events than when judging trivia questions. Notably, Ronis and Yates found that participants expressed judgments of complete certainty (judged probabilities of 0 or 1) on 25% of their responses to trivia questions, compared to 1.3% of responses to basketball games. Similarly, Olson and Budescu (1997) asked subjects to provide probabilistic judgments to either the outcomes of “spinners” or to almanac questions. For both linguistic and numerical judgments of probability, they found a striking difference in response distributions across the two different types of uncertainty. For almanac questions, responses spiked at 0, .5, and 1, with few judgments falling between these values. In contrast, the response distribution for spinner outcomes was considerably more uniform, with very infrequent use of the extreme points of the scale.

The foregoing studies provide indirect evidence for the notion that judgment extremity covaries with the nature of perceived uncertainty. In the studies that follow we provide more direct evidence. More specifically, we hypothesize that: (i) individuals reliably distinguish events in terms of their epistemicness (knowability) and aleatoriness (randomness), (ii) judgments will be more extreme for events that are perceived to be more epistemic (knowable) and more regressive for events that are perceived to be more aleatory (random), and (iii) such differences in judgment extremity can be attributed to differences in how evidence strength is mapped onto degrees of belief rather than differences in the evidence that is recruited for the events under consideration. We begin in Study 1 by providing initial evidence that judgment extremity varies systematically with perceived epistemicness and aleatoriness. We next provide a simple mathematical framework that will allow us to formally test our hypothesis about evidence sensitivity across different judgment domains (Studies 2A and 2B), within a single domain in which we manipulate relative epistemicness and aleatoriness (Study 3), and in a situation in which we prime participants to see a task as more epistemic or aleatory (Study 4). Following this exploration of judgment extremity, we examine judgment accuracy across all relevant studies. We consistently find that more extreme probability judgments associated with perceptions of greater epistemicness (knowability) entail a trade-off in different components of judgment accuracy. In particular, perceptions of greater epistemicness are associated with increased resolution of probability judgments (i.e., better discrimination) at the expense of decreased calibration (i.e., generally greater overconfidence). Moreover, we document that the observed pattern of judgment extremity can inform strategies for improving judgment accuracy in task environments for which question items are relatively easy versus difficult. Finally, we observe that the differences in evidence sensitivity under epistemic and aleatory uncertainty are amplified when judges rate themselves as more knowledgable concerning relevant events.

Study 1: Judgment Extremity Increases with Perceived Epistemicness

For Study 1, we recruited a sample of basketball fans to provide subjective probabilities for games in the first round of the 2014 NCAA men's college basketball tournament. We expected individuals to vary in their beliefs concerning the degree of epistemic (knowable) and aleatory (random) uncertainty involved in predicting basketball games, and that this variation would covary with the extremity of their judgments. In particular, individuals who view basketball outcomes as more knowable should provide more extreme judgments than individuals who view basketball outcomes as more random. Assigning subjective probabilities to basketball games in the NCAA tournament provides a clean initial 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 games that vary in team parity (e.g., a 1^{st} ranked team playing a 16^{th} ranked team, an 8^{th} ranked team playing a 9^{th} ranked team), thus allowing us to examine judgments that should vary across a wide range of probabilities.

Our sample consisted of 117 self-identified college basketball fans (60% female, mean age = 26 years, range: 18–63 years) who were recruited through an e-mail list, maintained by the UCLA Anderson School of Management, of individuals interested in receiving online survey announcements. Participants were paid a fixed amount in return 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 their knowledge of NCAA basketball for the current season (1 = *not at all knowledgeable*, 7 = *extremely knowledgeable*), and only those who reported knowledge above the midpoint of this scale were allowed to proceed to the study. Participants then provided probability judgments[1] for 30 games from the first round of the NCAA tournament.[2] For each trial, participants were reminded of each team's seeded ranking, and were asked the probability that a designated team would beat their opponent on a 0–100% scale. We randomized the order of trials, as well as the team designated as focal for a given trial[3] (i.e., whether participants judged $p(A \text{ defeats } B)$ or $p(B \text{ defeats } A)$).

In order to incentivize thoughtful responses, participants were told that some respondents would be selected at random and awarded a bonus of up to a $100, in proportion to their accuracy (based on their Brier score; see the supplemental materials for the full task instructions). Afterwards, three of the games were randomly sampled and participants rated the degree of epistemic and aleatory uncertainty associated with predicting a correct outcome. This was done using a 10-item epistemic-aleatory rating scale (EARS) that has been developed and validated elsewhere (Fox et al., 2014). The scale prompted participants to rate their agreement with a set of statements that measures both feelings of epistemic uncertainty (e.g., "determining the outcome to this question depends on knowledge or skill") and aleatory uncertainty (e.g., "the outcome to this question feels like it is determined by chance factors"). For the studies reported here, we reversed coded the aleatory items and averaged all items to form a single index (henceforth referred to as "epistemicness") with higher numbers indicating relatively greater epistemic uncertainty and lower numbers indicating greater aleatory uncertainty[4] (Cronbach's $\alpha = .74$).

Following the disclosure guidelines recommended by Simmons et al. (2011), we have provided the full list of measures for this study, as well as all subsequent studies, in the supplemental materials.

Study 1 Results

We conducted analyses at the trial-level using linear regression,[5] with participants treated as random effects and question items (i.e., individual basketball match-ups) as fixed effects. We

[1] In addition to probability judgments, we also elicited certainty equivalents for each basketball game. Participants evaluated prospects of $160 if a particular team won their game compared to a series of ascending sure payments (e.g., receive $50 for sure; see Fox, 1999, for a detailed description of this procedure). Data from the pricing task revealed considerable incoherence and we suspect that a large proportion of participants misunderstood the task instructions. For this reason, we excluded this portion of the study from the results.

[2] This included all first-round games excluding two play-in games that had yet to be played before we ran the study. Apart from the two play-in games (one 16^{th} seed and one 11^{th} seed), this left four each of every strength match up (1 vs. 16, 2 vs. 15, 3 vs. 14, and so forth).

[3] 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 .5 to 1. We chose the designated form for eliciting beliefs because it allows us to distinguish overextremity (the tendency to provide judgments that are too close to 0 or 1) from overprediction (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).

[4] In prior development of the EARS (Fox et al., 2014), 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 (knowable) uncertainty and decreasing aleatory (random) uncertainty — we treat them as a single dimension for simplicity by reverse-coding aleatory scale items. 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.

[5] Because we make ex ante predictions about the direction of the relationship between perceived epistemicness and judged probability, we report one-tailed tests throughout the paper.

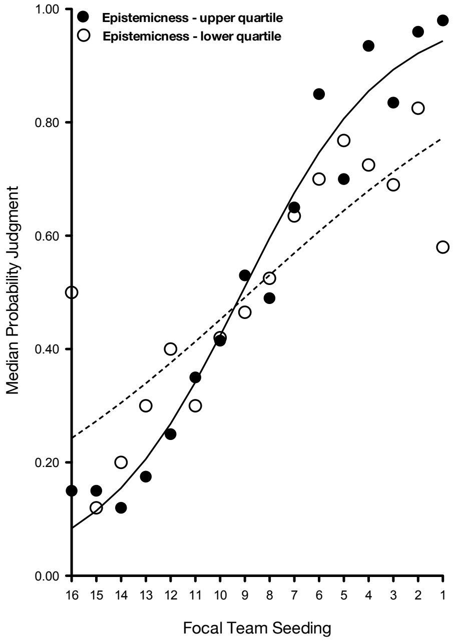

Figure 1: Study 1 Results. The x-axis represents the seeded ranking for the target team, and the y-axis represents judged probability. For each seeding, median judgments were calculated for the upper and lower quartile of responses in rated epistemicness. Lines represent the best fitting lines from a fractional response model.

analyzed judgment extremity by taking the absolute deviation in judged probability from $1/2$. As predicted, judgments were more extreme when epistemic uncertainty was especially salient; for every one-unit increase in perceptions of epistemicness, judgment extremity increased by an average of 4.4 percentage points ($b = .044$, SE $= .01$, $p < .001$). Furthermore, this pattern was true both when looking at judgments above and below one-half: perceptions of epistemicness were positively correlated with judgments above .50 ($b = .046$, SE $= .01$, $p < .001$), and negatively correlated with judgments below .50 ($b = -.027$, SE $= .01$, $p = .028$). Thus, we can rule out the possibility that our results are driven by a tendency toward greater extremity on only one end of the probability scale. We also ran a logistic regression on the likelihood of expressing a judgment of complete certainty (i.e., dummy coded probability judgments of either 0 or 1), and again found that perceptions of greater epistemicness were associated with more frequent extreme judgment, with an average marginal effect[6] of 12.5% ($p < .001$). Rows 2–5 of Table 1 report summary statistics for each quartile of rated epistemicness, and Figure 1 depicts the median judged probability as a function of team seeded ranking at the upper and lower quartiles of perceived epistemicness.

[6] 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 12.5 percentage point increase in complete certainty responses for every one-unit increase in epistemicness ratings.

Connecting Judgment Extremity to Evidence Sensitivity

Study 1 demonstrates that judged probabilities were more extreme for individuals who viewed basketball outcomes as entailing relatively more epistemic (knowable) uncertainty. In Studies 2–4, we examine how these differences in judgment extremity can arise from differences in *sensitivity to evidence*; that is, how people map the difference in assessed evidence for a focal event versus its complement onto judged probability. As discussed in the introduction we hypothesize that when people have an impression of more epistemic (knowable) uncertainty they become more sensitive to differences in evidence strength; small differences in the strength of evidence between competing hypotheses should translate into more extreme probability judgments. Conversely, heightened perceptions of aleatory (random) uncertainty should lead to diminished evidence sensitivity and relatively regressive judgments, holding evidence strength constant.

An analysis of the relative strength of evidence also allows us to examine differences in extremity across domains while controlling for parity of strength of hypotheses drawn from each domain. To illustrate the importance of this analysis, suppose we asked participants to judge the probability that various basketball teams and various football teams will win their upcoming games. More extreme probabilities for football than basketball could reflect differences in perceived epistemicness, but they could also reflect a belief that the selected football teams are more imbalanced than the selected basketball teams. Controlling for explicit ratings of evidence strength allows us to remove this potential confound and provides a common metric by which to compare sensitivity across domains.

Sensitivity to evidence strength can be formalized using support theory (Tversky and Koehler, 1994; Rottenstreich and Tversky, 1997), a formal model of judged probability. In support theory, probabilities are attached to *hypotheses*, or descriptions of events[7], with each hypothesis A generating a non-negative support value, $s(A)$. Support values can be thought of as representing 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 generated for a focal hypothesis relative to the support for its complement. That is, the probability that hypothesis A rather than the complementary hypothesis B obtains is given by

$$p(A, B) = \frac{s(A)}{s(A) + s(B)}. \tag{1}$$

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:

[7] 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 non-extensional). In the present paper we assume a canonical description of events, so this distinction will not be relevant.

Table 1: Epistemicness Ratings and Judgment Extremity in Studies 1–4

		Judgment Extremity			
	Epistemicness	MAD from	median	median	proportion
	M (*SD*)	$p = .50$	$p > .50$	$p < .50$	$p = 0$ or 1
Study 1					
4^{th} quartile	5.16 (0.46)	.32	.89	.16	.14
3^{rd} quartile	4.45 (0.12)	.28	.85	.22	.09
2^{nd} quartile	4.06 (0.10)	.23	.78	.30	.01
1^{st} quartile	3.10 (0.68)	.15	.70	.35	.01
Study 2A					
Geography	6.09 (1.13)	.28	.80	.10	.27
Population	5.90 (1.23)	.33	.90	.10	.16
Oceans	5.74 (1.21)	.36	.95	.10	.41
Crime	4.53 (1.55)	.31	.80	.20	.13
Housing	4.04 (1.52)	.20	.70	.30	.03
Temperature	3.19 (1.19)	.20	.75	.30	.02
Rain	3.14 (1.23)	.15	.62	.30	.01
Movies	3.09 (1.46)	.24	.80	.25	.09
Politics	2.95 (1.13)	.16	.60	.35	.05
Baseball	2.49 (1.28)	.12	.67	.30	.02
Football	2.49 (1.05)	.10	.65	.30	.00
Soccer	2.43 (1.17)	.11	.65	.40	.02
Study 2B					
Geography	6.01 (1.09)	.28	.90	.10	.33
Temperature	3.97 (1.21)	.23	.80	.20	.15
Basketball	3.33 (1.12)	.19	.70	.30	.10
Study 3					
Historic average task	4.93 (1.04)	.30	.80	.10	.20
Arbitrary day task	4.35 (1.08)	.25	.80	.20	.08
Study 4					
Pattern prediction condition	2.90 (1.47)	.20	.75	.25	.05
Random prediction condition	2.80 (1.60)	.17	.70	.30	.02

Note: MAD = mean absolute deviation.

$\hat{s}(A) \geq \hat{s}(B)$ 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.

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 evidence strength, or how evidence strength is mapped onto a probability judgment. Stated differently, even though judged probability should always increase whenever the balance of evidence favors a focal hypothesis, the rate at which it increases can vary — and the scaling constant represents this rate of change. This point can be seen more easily by converting probabilities into odds. Using Eq. (1), assuming all probabilities are positive, and defining $R(A, B)$ as the odds that A rather than B obtains, we get:

$$R(A, B) \equiv \frac{p(A, B)}{1 - p(A, B)} = \frac{s(A)}{s(B)} = \left[\frac{\hat{s}(A)}{\hat{s}(B)}\right]^k. \tag{2}$$

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 $\hat{s}(A)/\hat{s}(B)$ 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, we hypothesize that k should be larger when tasks are viewed as more epistemic (knowable),[8] and smaller when tasks are viewed as more aleatory (random).

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 Eq. (2):

$$\ln R(A, B) = k \ln \left[\frac{\hat{s}(A)}{\hat{s}(B)}\right]. \tag{3}$$

[8] A related observation 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 it is epistemic or aleatory — rather than sheer predictability.

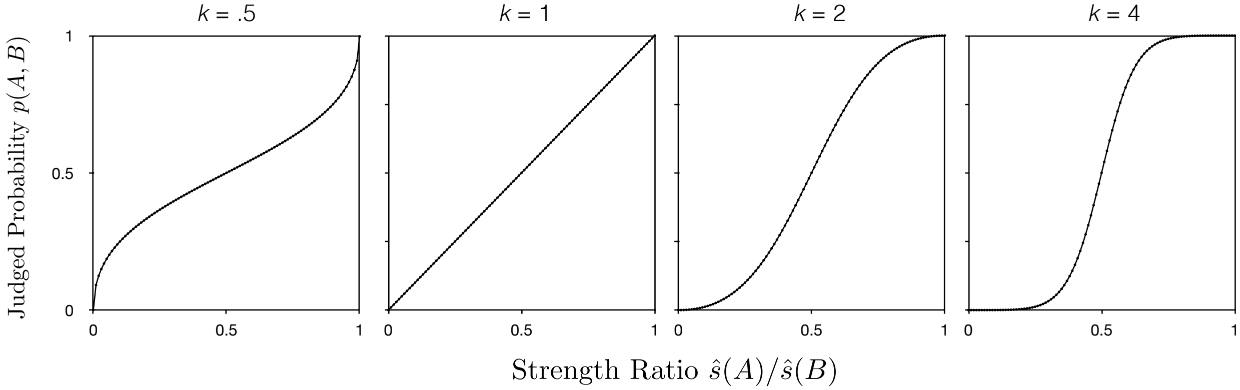

Figure 2: Examples of sensitivity to evidence strength (k).

Thus, using eq. (3) we can empirically estimate sensitivity to evidence strength by means of linear regression, 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 2A: Differences in Evidence Sensitivity Across Domains

In Study 2A we examined evidence sensitivity across a wide variety of domains, with the expectation that across-domain differences in evidence sensitivity would be correlated with across-domain differences in judged epistemicness. We recruited a sample of 206 participants from Amazon.com's Mechanical Turk labor market (MTurk) and paid them a fixed amount in return for their participation (56% female, mean age = 33 years, range: 18–80 years). One participant reported using outside sources (e.g., Wikipedia.org) to complete the task, and was dropped from the analysis. Participants first provided probability judgments to six questions that were randomly sampled from a pool of 12 questions, listed in Table 2, 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[9] for the two targets in each of their six previous estimates (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 (see the supplemental materials for all instructions). 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

[9] In studies 2–4 we excluded a small number of trials 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. Such responses are not directly interpretable, and also imply a misunderstanding of the task scale. This required us to drop 6% or less of all trials per study.

Table 2: Study 2A Stimulus Materials

Domain	Sample 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 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 Pittsburgh?
Politics	Assume that Barack Obama will face Mitt Romney in the 2012 presidential election. What is the probability 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 LA Dodgers on August 3rd. What is the probability that the Chicago Cubs will beat the LA Dodgers?
Movie sales	Consider two upcoming summer movies, The Amazing Spider-Man and The Dark Knight Rises. What is the probability that The Amazing Spider-Mane will gross more money on its opening weekend than The Dark Night Rises?
Real estate	Consider housing prices in Nashville and Atlanta. What is the probability a randomly-selected house in Nashville will be more expensive than a randomly-selected house in Atlanta?
Crime rates	Consider crime rates in Detroit and Columbus. What is the probability the number of violent crimes per capita this year will be higher in Detroit than Columbus?
Geography	Consider the geographic size (in sq. 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?
Ocean size	Consider the size (in sq. miles) of the Atlantic Ocean and Indian Ocean. What is the probability that the Atlantic Ocean is larger than the Indian Ocean?

> 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.

Finally, participants were again shown each of the six events whose probabilities they had previously assessed and rated each event on an abridged 4-item epistemicness scale (Cronbach's α ranged from .60 to .87 across domains, with an average score of .75; see supplemental materials for scale items).

Study 2A Results

First we examined for judgment extremity. Conceptually replicating Study 1, we found more extreme judgments for domains higher in perceived epistemicness (knowability). Table 1 lists the

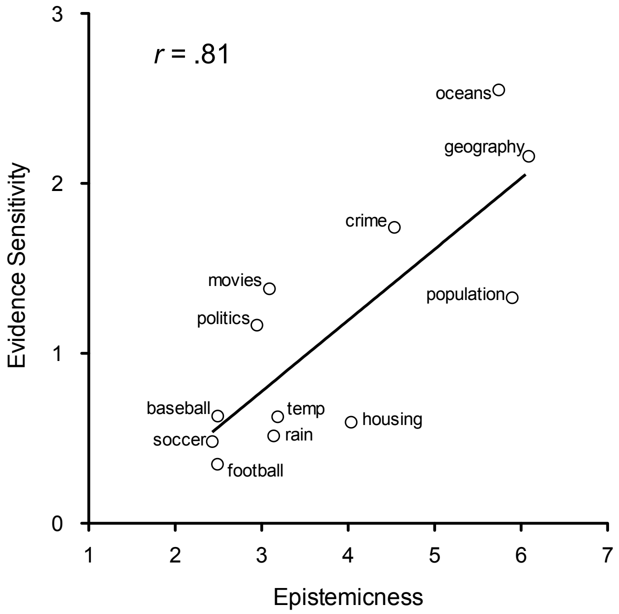

Figure 3: Relationship between evidence sensitivity (k) and rated epistemicness (Study 2A).

average epistemicness rating by domain (see rows 7–18), along with indices of judgment extremity. Averaging over responses by domain, the correlation between rated epistemicness and mean absolute deviation was positive and substantial ($r = .90$, $p < .001$). At the individual level, there was an average correlation of .43 between participants' epistemicness ratings and mean absolute deviation in judgments. We also obtained similar results when restricting the analysis to judgments above .50, below .50, or to judgments of 0 and 1 (p-values $< .001$).

Next, we estimated sensitivity to differences in evidence strength. As discussed above, an analysis of evidence sensitivity allows us to more rigorously account for the nature of events in our sample and their distribution (i.e., an analysis of evidence sensitivity controls for domain-level differences in item parity). This was done at the trial level by regressing the log odds of judged probabilities[10] onto the log strength ratio of the hypotheses under consideration, as suggested by eq. (3)

$$\ln\left[\frac{p(A,B)}{1-p(A,B)}\right] = \alpha + \beta \ln\left[\frac{\hat{s}(A)}{\hat{s}(B)}\right] + \epsilon. \tag{4}$$

The intercept α can be interpreted as an index of response bias,[11] and the coefficient β can

[10] Because the regression analysis required transforming probability judgments into log odds, responses of 0 or 1 were compressed by substituting them with $.5/N$ and $[N - .5]/N$, respectively, for sample size N, as suggested by Smithson and Verkuilen (2006).

[11] While not the primary focus of the current research, a key assumption of support theory is *binary complementarity*, which implies that judgments to two-alternative outcomes are additive (i.e., $p(A,B) + p(B,A) = 1$). Thus, according to binary complementarity we should not see any appreciable degree of response bias (i.e., the intercept should not differ substantially from 0). Indeed, replicating previous research we found that the intercept term of all studies, with the exception of Study 3, did not reliably differ from 0. Direct tests of binary complementarity are provided in the supplementary materials.

Table 3: Sample stimuli across domains (Study 2B)

Domain	Question
Basketball	Suppose that the Los Angeles Clippers play the Boston Celtics in the NBA finals. What do you think is the probability that the Los Angeles Clippers will win?
Temperature	Consider a day picked at random next year in Los Angeles and Minneapolis. What do you think is the probability that it will be warmer in Los Angeles that day?
Geography	Consider the geographical size of Wisconsin and Georgia. What do you think is the probability that Wisconsin is the larger state?

be interpreted as an index of evidence sensitivity, and was estimated for each of the 12 judgment domains. Estimates of β were then correlated with the average epistemicness rating for each domain. Figure 3 depicts domain estimates of k plotted against corresponding ratings of epistemicness, and reveals a substantial positive correlation (r = .81). That is, judgments were most sensitive to differences in evidence strength for domains entailing more epistemic (knowable) uncertainty. Using the predicted point estimates from the model, we would expect to see a 4.5-fold increase in evidence sensitivity when going from the domain lowest in epistemic uncertainty to the domain highest in epistemic uncertainty (i.e., a larger effect than going from the second panel in Figure 2 to the fourth panel).

Study 2B: Differences in Evidence Sensitivity Within Domains

Study 2A demonstrated that across-domain differences in evidence sensitivity varied as a function of perceived epistemicness. In Study 2B we sought to complement this finding by also exploring within-domain differences in epistemicness. To do so, we selected three judgment domains expected to span the range of perceived epistemicness — geography questions, weather estimates, and upcoming NBA basketball games. Based on the findings of Study 2A, we expected that questions about U.S. geography would be viewed as primarily epistemic, NBA basketball games would be viewed as primarily aleatory, and weather-related events would be viewed between these extremes.

Study 2B Methods

We recruited a sample of 37 self-identified basketball fans[12] from MTurk (19% female, mean age = 33 years, range: 19–59 years) who participated in return for a small cash payment plus entry into a drawing to receive an NBA basketball jersey of their choice. One participant reported using outside sources while completing the task, and was dropped from the analysis.

First, participants provided probability judgments to 16 two-alternative questions for each of three domains: (i) outcomes of upcoming NBA playoff games, (ii) outcomes of upcoming daytime

[12] We planned to sample 50 basketball fans, but successfully recruited only 37 participants before the start of the NBA playoffs. Note that the design of Study 2B was entirely within-subject, so the sample still provided reasonable statistical power.

high temperatures, and (iii) general knowledge questions about the relative geographic sizes of different U.S. states (see Table 3 for sample questions). As before, one alternative was designated as focal for each question (counterbalanced between participants) and participants estimated the probability (from 0-100%) that the focal team would win their game (basketball), the focal city would register a higher daytime high temperature (weather), or the focal state is geographically larger (geography). For each domain we constructed two sets of four targets (e.g., Atlanta, Buffalo, Los Angeles, Memphis; Miami, Minneapolis, New Orleans, San Francisco) and defined events by factorially pairing one target from each set (e.g., Atlanta-Miami) so that we had sixteen possible pairings for each domain. The ordering of judgment domains and questions within domains were randomized, with the constraint that all judgments within a domain were completed before advancing to the next block. Next, participants were provided with a list of the targets from each domain and were asked to assess their relative strength (strength of teams, warmth of cities, size of states) following the same protocol as in Study 2A. In the final phase of the study participants rated each domain for its degree of epistemicness. A single trial was selected at random from each domain and participants rated the event using the 10-item EARS scale used in Study 1 (Cronbach's α ranged from .83 to .88 across domains).

Study 2B Results

Column 2 of Table 1 provides average epistemicness ratings for the three domains (see rows 19–21). Our NBA basketball fans rated basketball as the least epistemic domain, followed by temperature estimates, and then by geography questions — all consistent with the pattern that we had observed for sports versus weather versus geography in Study 2A. All means were reliably different from one another (p-values $< .001$). More importantly, and consistent with our hypothesis, both judgment extremity and evidence sensitivity followed the same rank-ordering as epistemicness ratings — smallest for basketball, intermediate for city temperature, and highest for the state geography questions.

As expected, mean absolute deviations from 1/2 were least extreme for basketball games, intermediate for temperature estimates, and most extreme for geography questions. All means were reliably different from one another (p-values $< .01$). We obtained similar results when restricting the analysis to judgments above .50 ($p < .01$ for all pairwise comparisons), below .50 ($p < .01$ for all pairwise comparisons), or to judgments of 0 and 1 ($p < .01$ for all pairwise comparisons except temperature vs. basketball, $p = .102$; see Table 1 for summary statistics).

Next, we examined estimates of sensitivity to evidence strength. This was first done at the trial-level by regressing strength ratings onto probability judgments (converted to log odds), with participants treated as random effects and judgment domains as fixed effects. The first column in Table 4 (Model I) displays the average marginal effects (i.e., estimates of k) for each domain. As expected, estimates were smallest for basketball predictions, intermediate for temperature estimates, and largest for geography questions. We also analyzed the data at the subject-level (by running separate regressions for each participant per domain; Model II) and at the item-level (by taking the

Table 4: Study 2B Estimates of Evidence Sensitivity

	Model I	Model II	Model III
Geography	2.20^a (0.10)	3.21^a (0.37)	4.90^a (0.22)
Temperature	1.95^a (0.12)	3.00^a (0.37)	2.63^b (0.18)
Basketball	1.34^b (0.12)	2.03^b (0.37)	1.57^c (0.31)
Unit of Analysis	trials	subjects	items
No. of observations	1,716	108	96
No. of groups	36	36	52
R^2	.331	.052	.892

Note: Estimates represent coefficients from linear regression. For all models, observations are nested within groups (as random effects) and task domains are treated as fixed effects. Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates ($p < .05$).

median response for each possible item-pairing; Model III). Again, we found the same qualitative pattern of results.

Thus far we have found that the rank-ordering of domains in perceived epistemicness corresponded to the rank-ordering in sensitivity to evidence strength. Next we more directly examined the relationship between rated epistemicness and evidence sensitivity. At the trial-level of analysis (i.e., estimates derived in Model I of Table 4), this would imply a positive interaction between strength ratings and perceived epistemicness — the slope on strength ratings, which represents an estimate of k, should increase as perceived epistemicness increases. Indeed, we found a reliable and positive interaction effect ($b_{intx} = 0.35$, SE = 0.04, $p < .001$). Based on the regression coefficients, k would be expected to increase from 1.24 to 2.38 when going from one standard deviation below to one standard deviation above the mean in perceived epistemicness.

Study 3: Holding Strength Measurement Constant

Studies 2A and 2B suggest that domains higher in rated epistemic (knowable) uncertainty are associated with greater sensitivity to evidence strength. One limitation of these studies is that different domains require different measures of evidence strength. 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 in Study 2A 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 Studies 2A and 2B were, for whatever reason, more appropriate proxies of hypothetical support. Note that this cannot explain readily our finding that individual differences in perceived epistemicness (knowability) were related to differences in evidence sensitivity. Nevertheless, it would be desirable to replicate such an analysis for events that are matched in

Table 5: Study 3 Sample Questions

Task Format	Sample Question
Historic Average	What is the probability that the 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?

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 in their extremity of judgment.

Toward that end, in Study 3 we asked participants to estimate the probability that one of two U.S. cities had a higher daytime high temperature. Crucially, participants compared cities based on: (a) their historic averages, and (b) an arbitrarily-selected day over the same time period. Naturally, global impressions of evidence strength — in this case, that one city is "warmer" than another — will be more diagnostic for historic averages than for single-days, since there is greater fluctuation in temperatures over individual days than over an average of a collection of days. The more interesting question is whether individual variation in impressions of relative epistemicness corresponds with individual variation in judgment extremity (while also holding the strength elicitation procedure fixed across the two task conditions).

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, 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 (blocks were presented in a randomized order). For trials in the *historic 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. Table 5 provides sample questions. Each block consisted of 15 trials (by forming all pairwise comparisons between 6 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 the same 10-item EARS scale as in the previous studies. After responding to both judgment blocks, participants provided strength ratings for the 6 cities in a manner similar to Studies 2A and 2B.

Table 6: Study 3 Estimates of Evidence Sensitivity

	Model I	Model II	Model III
Historical average	2.04^a (0.04)	3.39^a (0.25)	3.15^a (0.13)
Arbitrary day	1.44^b (0.04)	2.25^b (0.16)	1.82^b (0.13)
Unit of Analysis	trials	subjects	items
No. of observations	5,762	392	60
No. of groups	196	196	30
R^2	.422	.063	.930

Note: Estimates represent coefficients from linear regression. For all models, observations are nested within groups (as random effects) and question items are treated as fixed effects. Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates ($p < .05$).

Study 3 Results

Not suprisingly, participants rated the historic average task as entailing greater epistemicness than the arbitrary day task (means were 4.93 and 4.35, respectively; $p < .001$). Moreover, as displayed in rows 23–24 of Table 1, we found greater judgment extremity in the historic average task than in the arbitrary day task; the mean absolute deviation from $1/2$ was on average 4.8 percentage points higher when participants provided judgments about historic averages compared to arbitrarily-selected days ($b = .048$, SE = .003, $p < .001$). We also saw greater judgment extremity in the historic average task when restricting the analysis to judgments above .50 ($b = .040$, SE = .004, $p < .001$), to judgments below .50 ($b = -.041$, SE = .004, $p < .001$), or to judgments of 0 and 1 (20% vs 8%, $p < .001$). Third, conceptually replicating the results from Studies 2A and 2B, we observed greater evidence sensitivity at the trial-, subject-, and item-levels when responding to estimates of historic averages than arbitrarily selected days (see Table 6).

Because both judged probabilities and perceptions of epistemicness were measured within-subjects, we were able to assess the direct relationship between judgment extremity and perceived epistemicness. First, we examined the relationship between epistemicness and evidence sensitivity at the trial-level by probing for a positive interaction between strength ratings and perceived epistemicness — the coefficient of strength ratings, which is an estimate of k, should increase as perceived epistemicness increases. As expected, we found a reliable and positive interaction effect ($b_{intx} = 0.29$, SE = 0.06, $p < .001$). Based on the regression coefficients, k would be expected to increase from 1.35 to 1.99 when going from one standard deviation below to one standard deviation above the mean in perceived epistemicness.

Recall that the main aim of Study 3 was to identify whether individual variation in impressions of relative epistemicness could explain individual variation in judgment extremity. To do this, we recovered estimates of k for each participant (by running separate regressions on each participant's judgments) separately for the historic average task and the arbitrary day task. We then examined the correlation between the difference in a participant's evidence sensitivity between the two tasks ($k_{average} - k_{arbitrary}$) and the difference between the participant's mean epistemicness ratings between the two tasks (epistemicness$_{average}$ − epistemicness$_{arbitrary}$). As expected, we find a positive and significant correlation ($r = .24, p < .001$). That is, participants who showed the largest differences in their perceptions of epistemicness across the two tasks also tended to exhibit the largest differences in sensitivity to evidence strength.

Study 4: Priming Epistemic and Aleatory Uncertainty

Study 3 demonstrated that tasks rated higher in epistemic uncertainty were associated with more extreme judgments and greater sensitivity to evidence strength, even when the elicitation procedure for assigning strength ratings was held constant across conditions. In Study 4, we directly manipulate perceptions of epistemicness while holding all features of the judgment task constant, thereby providing a stronger test of the hypothesis that perceptions of epistemicness causally influence the mapping of evidence strength onto judgment. To manipulate perceptions of epistemicness/aleatoriness, we asked participants 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 sub-optimal (because choosing the higher expected value option on every trial will maximize a participant's 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 to 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 UCLA subject pool, who were each paid a fixed amount for their participation in a laboratory study (82% female, mean age = 20 years, 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. After 10 practice trials, participants completed 168 trials divided into two blocks of 84 trials. In one block participants viewed trials that were generated randomly, while 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 Xs and Os 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.

Our key manipulation was to vary how this first phase of the study was described to participants (see the supplementary materials for the full set of instructions). In the *pattern prediction* 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.

In the second phase of the study participants provided 28 probability judgments to upcoming weather-related events in eight U.S. cities. For each trial, participants were presented with two cities (sampled from a pool of eight cities), with one city designated as focal. Participants indicated the probability that a given city would have the higher daytime high temperature on the following July 1st. 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 fourth phase of the study, participants were presented with three randomly-selected trials from phase two, and rated each question on the 10-item epistemic-aleatory rating scale (EARS) used in the previous studies. We averaged these three trials to form an index of perceptions of epistemicness for the judgment task (average Cronbach's $\alpha = .75$).

Study 4 Results

As a manipulation check, we examined average epistemicness scores for each task. Questions were viewed as entailing more epistemic (knowable) uncertainty when participants were prompted to seek patterns than when they were prompted to guess, although this difference was not statistically significant (means were 4.16 and 4.04, respectively; $p = .22$). However, when we separately examined epistemic and aleatory items from the scale we noted that the correlation between these two

Table 7: Study 4 Estimates of Evidence Sensitivity

	Model I	Model II	Model III
Pattern Recognition task	1.35^a (0.06)	2.03^a (0.23)	2.12^a (0.11)
Random Guessing task	1.03^b (0.05)	1.20^b (0.22)	1.10^b (0.12)
Unit of Analysis	trials	subjects	items
No. of observations	2,795	100	112
No. of groups	100	100	56
R^2	.279	.063	.804

Note: Estimates represent coefficients from linear regression. For all models, observations are nested within groups (as random effects) and question items are treated as fixed effects. Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates (p < .05).

indices was particularly weak[13] ($r = -.16$). Thus, we proceeded to analyze each of these subscales separately, finding no reliable difference in ratings on the epistemic subscale (means were 4.70 and 4.78, respectively; $p = .67$) but a significant effect in the expected direction on the aleatory subscale (means were 4.63 and 5.11, respectively; $p = .035$). That is, participants viewed questions as higher in aleatory uncertainty for the random prediction prime than for the pattern prediction prime. Our manipulation, it appears, was more successful at priming aleatory (random) uncertainty than epistemic (knowable) uncertainty. While these tests suggest that the persistent impact of our manipulation was modest by the time participants reached the last phase of our study (i.e., the manipulation check), we note that prior research has found that the impact of similar primes are often short-lived (Simon et al., 2008).

More importantly, and as predicted, probability judgments were more extreme when participants were primed with pattern prediction than random prediction (results are summarized in rows 26–27 of Table 1). Using mean absolute deviation from 1/2, judgments were on average 2.7 percentage points more extreme for the pattern prediction task than the random prediction task ($b = .027$, SE = .016, $p = .039$). Furthermore, we see greater judgment extremity when restricting the analysis to judgments above .50 ($b = .029$, SE = 0.015, $p = .028$), to judgments below .50 ($b = -.033$, SE = .017, $p = .024$), or to judgments of 0 and 1 (5% vs 2%, $p = .053$).

Most importantly, we observed greater sensitivity to evidence strength in the pattern prediction task than in the random prediction task. As before, we analyzed evidence sensitivity at the trial-level, subject-level, and item-level (Models I–III in Table 7), and for all three models estimated k was significantly higher in the pattern prediction task (all p-values < .05). For example, when comparing conditions at the trial-level (Model I), we found a 1.3-fold increase in sensitivity to evidence strength between the two task conditions.

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

[13] For comparison, the correlation between the epistemic and aleatory items was stronger than that observed in Study 4 for all studies with the exception of Study 1. As mentioned earlier, however, results for all studies held when analyzing the data separately by epistemic and aleatory subscales.

perceived epistemicness — the coefficient of strength ratings, which is an estimate of k, should increase as perceived epistemicness increases. As expected, we found a positive interaction effect ($b_{intx} = 0.16$, SE = 0.10, $p = .057$). Based on the regression coefficients, k would be expected to increase from 0.91 to 1.17 when going from one standard deviation below to one standard deviation above the mean in perceived epistemicness. We also note that when restricting the interaction to the subset of aleatory items (which were more strongly affected by the prime according to our manipulation check), we see an even stronger interaction effect ($p = .009$).

General Discussion

The current research suggests that judgment is more extreme under uncertainty that is perceived to be more epistemic (knowable) and less aleatory (random). We observed this pattern within a single domain (Study 1), across distinct judgment domains (Studies 2A and 2B), when measures of evidence strength were matched across tasks (Studies 3 and 4), and when participants were experimentally primed to focus on epistemic or aleatory uncertainty (Study 4). These findings suggest that the perceived nature of the uncertainty underlying a judgment task can affect the extremity of one's beliefs. More broadly, this research suggests 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

The question arises as to what perceptions of epistemic and aleatory uncertainty tell us about judgment accuracy. At first glance, one might surmise that greater perceived epistemicness (knowability) should be associated with lower accuracy, given that individuals are generally prone to overconfidence. To examine this, we calculated accuracy scores for the three studies that were amenable to an analysis of judgment accuracy[14] (Studies 1, 2A, 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 f_i be the forecasted probability of event o on the ith occasion. The mean probability score is given by

[14] Study 2B was not included in the analysis because judgments of basketball games — which comprised one-third of the stimulus items in the study — were based on possible match-ups that were not all realized (e.g., "Suppose the San Antonio Spurs play the Philadelphia 76ers in the NBA finals ..."). Furthermore, questions about city temperatures involved estimating upcoming temperatures for days selected at random, 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 to general knowledge 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). For similar reasons, Study 3 was also excluded from the analysis.

$$\overline{PS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{5}$$

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, 2A, and 4 we regressed probability scores onto epistemicness ratings. We conducted analyses at the trial-level using a fractional logit model (which accommodates responses that are bounded between 0 and 1; Papke and Wooldridge, 1996), with participants treated as random effects and question items as fixed effects. Surprisingly, in all three studies we found no reliable increase or decrease in accuracy as a function of judged epistemicness. Column 2 of Table 8 provides the regression coefficients on $\overline{PS}$, and Table 9 reports summary statistics for each quartile of rated epistemicness in each study. As indicated in Table 8, we also failed to find reliable differences in the proportion of correct judgments[15] (i.e., judgment performance) as a function of epistemicness.

These results may first seem puzzling in light of our robust findings concerning judgment extremity. Apparently, perceptions of epistemicness are associated with more extreme probability judgments, but not with a reliable improvement in actual performance predicting outcomes as measured by Brier scores. The puzzle is resolved when we partition probability scores into interpretable components. Following Murphy (1973), we decompose probability scores as follows:

$$\begin{aligned}\overline{PS} &= \bar{o}(1-\bar{o}) + \left(\frac{1}{N}\right)\sum_{j=1}^{J} n_j(f_j - \bar{o}_j)^2 - \left(\frac{1}{N}\right)\sum_{j=1}^{J} N_j(\bar{o}_j - \bar{o})^2 \\ &= V + C - R,\end{aligned} \tag{6}$$

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 an event in that class, and $\bar{o}$ is the overall relative frequency of the event. For our analyses, judged probabilities were partitioned into tenths (i.e., judgments were separated into bins of 0–.10, .11–.20, and so forth).

The first term on the right-hand side of eq. (6) represents outcome variance (V), or the degree that the outcome varies from trial to trial. Outcome variance is usually interpreted as an indicator of task difficulty, and therefore 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. For example, a forecaster is considered well-calibrated if she assigns a judged probability of .20 to events that occur on 20 percent of occasions, a probability of .40 to events that occur on 40 percent of occasions, and so forth. The third term represents judgment resolution (R),

[15] Proportion correct was calculated at the trial-level by assigning a score of 1 when participants provided a judged probability above .50 and the event occurred, or a probability below .50 and the event failed to occur. Participants were assigned a score of 0 when judged probability was below .50 and the event occurred, or above .50 and the event failed to occur. For responses of .50, we randomly assigned responses as correct or incorrect (see Ronis and Yates, 1987).

Table 8: Average marginal effects of epistemicness on components of judgment accuracy

	Brier Score ($\overline{PS}$)	Proportion Correct	Calibration (C)	Resolution (R)
Study 1	.002	.008	.008***	.004***
	(.010)	(.024)	(.002)	(.001)
Study 2A	.008	.000	.003***	.002*
	(.005)	(.010)	(.000)	(.001)
Study 4	.017	.030	.003***	.006**
	(.014)	(.029)	(.001)	(.002)

*Note: Estimates represent average marginal effects from fractional logit regressions. Trials are nested within participants (treated as a random effect) with question items as a fixed effect. For Brier and Calibration Scores, positive coefficients are associated with decreased accuracy. For Resolution, positive coefficients are associated with increased accuracy. Standard errors in parenthesis; ***, **, * indicate significance at .001, .01, and .05 level, respectively.*

Table 9: Average Brier ($\overline{PS}$), Calibration (C), and Resolution (R) scores as a function of Judged Epistemicness

	Study 1			Study 2A			Study 4		
Epistemicness	$\overline{PS}$	C	R	$\overline{PS}$	C	R	$\overline{PS}$	C	R
4^{th} quartile	.271	.038	.026	.239	.031	.044	.217	.012	.040
3^{rd} quartile	.243	.030	.023	.254	.016	.038	.241	.006	.026
2^{nd} quartile	.231	.021	.017	.212	.011	.031	.213	.007	.027
1^{st} quartile	.274	.012	.011	.228	.008	.023	.215	.005	.023

Note: Epistemic quartiles are ordered from high (4^{th} quartile) to low (1^{st} quartile). For Brier and Calibration scores, lower numbers indicate greater accuracy. For Resolution, higher numbers indicate greater accuracy.

or the degree to which a forecaster reliably discriminates between events that do and do not occur. While calibration provides a measure of how close a judgment is to the truth, resolution provides a measure of the information contained in a forecast. For example, a forecaster who consistently provides a judged probability of 50% to a series of two-alternative outcomes (e.g., predicting which team will win their baseball game throughout the upcoming MLB season) will be well calibrated but undiscriminating in his judgment (that is, exhibit poor resolution). Note that superior performance is represented by low scores on C and high scores on R.

Returning to the previous results, we decomposed $\overline{PS}$ into calibration and resolution scores. As before, we conducted analyses at the trial-level using a fractional logit model, with participants treated as random effects and question items as fixed effects. The regression coefficients are displayed in columns 3–4 of Table 8, and are summarized in Table 9. For all three studies, a consistent pattern of results 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 (knowability) reduced accuracy by reducing calibration: participants were generally overconfident, and this tendency was exacerbated by more extreme judgments. On the other hand, heightened perceptions of epistemicness improved accuracy by improving resolution, as participants made fuller use of the probability scale. 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. That is, participants who saw events as more knowable and less random were both more *and* less accurate than participants who saw events as less knowable and more random, depending on the type of accuracy.

Epistemicness, Overconfidence, and Task Difficulty

Our findings have distinct implications for task conditions that lead to systematic overconfidence versus systematic underconfidence. One well-established finding is the tendency for participants to be overconfident for difficult tasks and underconfident for easy tasks (the "hard-easy" effect; e.g., Lichtenstein and Fischhoff, 1977; Soll, 1996). For difficult questions that few participants answer correctly, individuals tend to display overconfidence; this pattern is attenuated and may ultimately reverse to underconfidence as the level of task difficulty decreases.

If perceptions of epistemicness do not affect task performance but influence judgment extremity (as shown in the previous section), then we should expect perceptions of more epistemic versus aleatory uncertainty to improve accuracy under different task conditions. 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) while the regressiveness associated with low epistemicness should amplify underconfidence (diminish accuracy). Thus, we expect judgment accuracy 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 mean probability scores ($\overline{PS}$) onto item difficulty (operationalized as the proportion of correct responses per question), perceptions of epistemicness, and the interaction term. In all analyses we used a fractional logit model with participants treated as random effects. 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 $\overline{PS}$ as the proportion of correct responses increases, reflecting the fact that accuracy improves as task questions become less difficult. More importantly, in all three cases we found a reliable interaction effect that conforms

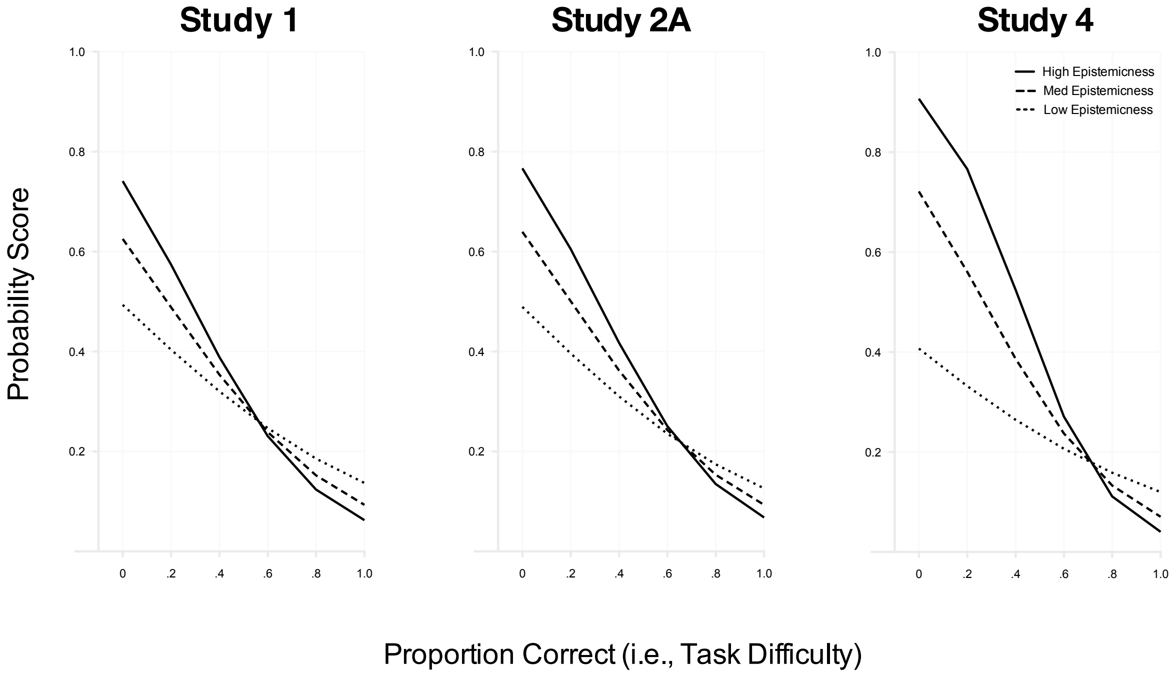

Figure 4: Accuracy as a function of task difficulty and judged epistemicness.

to the expected pattern of results (p-values < .001). Participants high in perceived epistemicness tended to be relatively less accurate (higher $\overline{PS}$) for difficult questions but more accurate (lower $\overline{PS}$) for easy questions, while participants low in perceived epistemicness tended to display the reverse pattern of results. Thus, heightened perceptions of epistemicness appeared to improve accuracy under some task conditions (i.e., easy questions) but diminished accuracy under other conditions (i.e., difficult questions), as predicted.

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 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 — 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 (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 that 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

Recently, 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 (as 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.

Studies 3 and 4 allowed for the most direct test of the random support model, as these studies held the strength elicitation format constant across conditions. For both studies 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 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 ranged from .301 to 1.00 and the observed R^2 from every test was always less than .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 knowledgable (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

$$R(A, B) = \left[\frac{n_A}{n_B}\right]^{1-\lambda} \left[\frac{\hat{s}(A)}{\hat{s}(B)}\right]^{k'\lambda}. \tag{7}$$

The second expression on the right-hand side of eq. (7) represents the balance of support as measured by raw strength ratings, akin to the original support theory formulation presented in eq. (2). 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. 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 regress towards 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 knowledgable, 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 eq. (2) has now been partitioned 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: evidence sensitivity as a function of perceived epistemicness should be amplified when participants are more knowledgable (i.e., when they rely less on the ignorance prior), and should be attenuated when participants are less knowledgable (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[16] for each judgment domain. For studies in which we measured sensitivity to evidence strength and asked participants to rate their knowledge separately for each domain or task (Studies 2A, 2B, 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

[16] For Studies 2A and 4, knowledge was assessed on a 11-point scale from 0 (*not knowledgable at all*) to 10 (*very knowledgable*). For Study 2B we assed knowledge in a similar manner but using a 100-point scale, which we subsequently transformed (by multiplying responses by .1) for purposes of comparison with Studies 2A and 4.

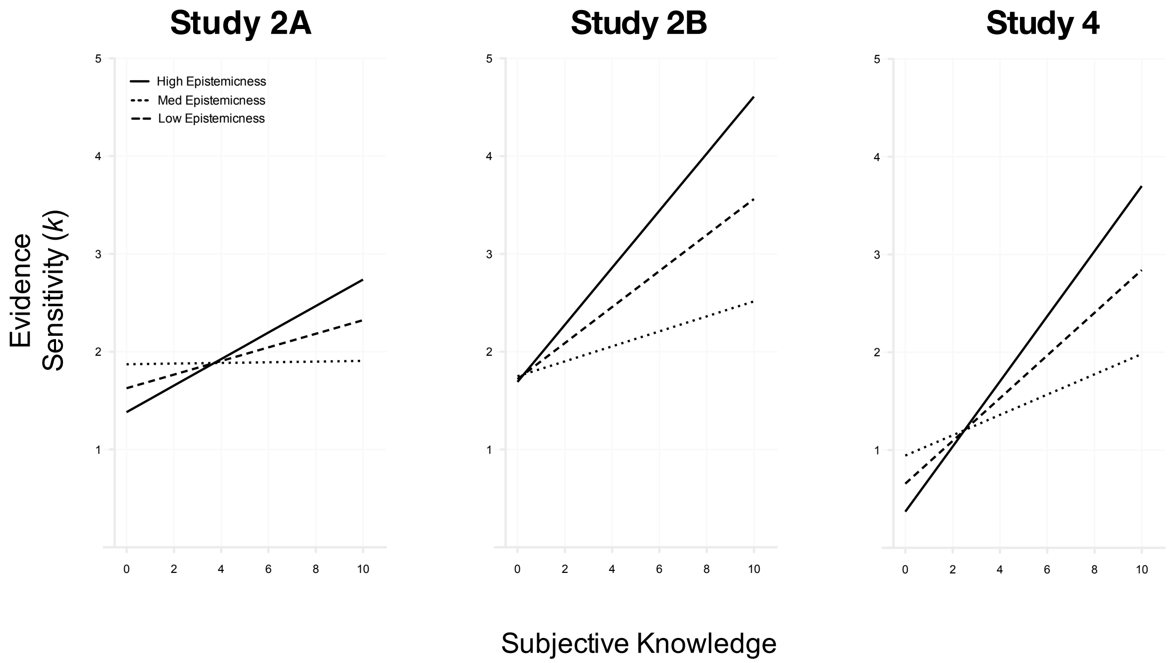

Figure 5: Sensitivity to evidence strength as a function of subjective knowledge and judged epistemicness.

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. That is, differences in sensitivity to evidence strength for high and low perceived epistemicness were most pronounced when knowledge was high. The interaction term between rated epistemicness and task knowledge was statistically significant in two of the three studies (p-values were .003, .082, and .007 for Studies 2A, 2B, and 4, respectively).

While 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 this prediction by independently manipulating the ignorance 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).

Conclusion

Experts and laypeople confront uncertainty on a near-constant basis. Whether evaluating an investment, engaging in geopolitical forecasting, or assessing the value of a potential corporate initiative, individuals must evaluate to varying degrees events that can be construed as knowable or random. In this paper we have documented a general tendency for judgments to be more extreme under heightened perceptions of epistemic (knowable) uncertainty and less extreme under heightened perceptions of aleatory (random) uncertainty. 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, may inform procedures and elicitation techniques for improving judgment accuracy.

References

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

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes*, 90(1):87–110.

Carlson, B. W. (1993). The accuracy of future forecasts and past judgments. *Organizational Behavior and Human Decision Processes*, 54(2):245–276.

Fischhoff, B. and Beyth, R. (1975). I knew it would happen: Remembered probabilities of oncefuture things. *Organizational Behavior and Human Performance*, 13(1):1–16.

Fischhoff, B. and Bruine De Bruin, W. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, 12(2):149–163.

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

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

Fox, C. R. and Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14(3):195–200.

Fox, C. R., Tannenbaum, D., and Ülkümen, G. (2014). The empirical case for distinguishing two dimensions of subjective uncertainty. Unpublished manuscript. University of California, Los Angeles.

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

Fox, C. R. and Ülkümen, Gülden, G. (2011). Distinguishing two dimensions of uncertainty. In Brun, W., Keren, G., Kirkebøen, G., and Montgomery, H., editors, *Perspectives on Thinking, Judging, and Decision Making: A tribute to Karl Halvor Teigen*, pages 21–35. Universitetsforlaget, Oslo.

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

Goodnow, J. J. (1955). Determinants of choice-distribution in two-choice situations. *The American Journal of Psychology*, 68(1):106–116.

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

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

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4):719–731.

Howell, W. C. and Kerkar, S. P. (1982). A test of task influences in uncertainty measurement. *Organizational Behavior and Human Performance*, 30(3):365–390.

Iachine, I., Petersen, H. C., and Kyvik, K. O. (2010). Robust tests for the equality of variances for clustered data. *Journal of Statistical Computation and Simulation*, 80(4):365–377.

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

Koehler, D. J. (1996). A strength model of probability judgments for tournaments. *Organizational behavior and human decision processes*, 66(1):16–21.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):107–18.

Levene, H. (1960). Robust tests for equality of variances. In Olkin, I., editor, *Contributions to Probability and Statistics*, volume 2, pages 278–292. Stanford University Press, Palo Alto, California.

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

Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20(2):159–183.

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

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600.

Olson, M. J. and Budescu, D. V. (1997). Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making*, 10(2):117–131.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., and Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development*, 77(6):1642–1655.

Ronis, D. L. and Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, 40(2):193–218.

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

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

See, K. E., Fox, C. R., and Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(6):1385–1402.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Simon, D., Krawczyk, D. C., Bleicher, A., and Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making*, 21(1):1–14.

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

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65(2):117–137.

Tversky, A. and Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4):547–567.

Ülkümen, Gülden, G., Fox, C. R., and Malle, B. F. (2014). Two faces of subjectives uncertainty: Cues from natural language use. University of Southern California.

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

Volz, K. G., Schubotz, R. I., and Von Cramon, D. Y. (2004). Why am i unsure? internal and external attributions of uncertainty dissociated by fmri. *Neuroimage*, 21(3):848–857.

Volz, K. G., Schubotz, R. I., and Von Cramon, D. Y. (2005). Variants of uncertainty in decision-making and their neural correlates. *Brain Research Bulletin*, 67(5):403–412.

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

Wolford, G., Newman, S. E., Miller, M. B., and Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, 58(4):221–228.

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

Wright, G. and Ayton, P. (1987). Task influences on judgemental forecasting. *Scandinavian journal of psychology*, 28(2):115–127.

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

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30(1):132–156.