# Epistemic versus Aleatory Judgment Under Uncertainty

## Introduction

Consider the following two cases:

1. Allie is playing Bingo at her local social hall. She is just one lucky number away from winning the game, but so are three of her friends. Allie is *uncertain* whether she will win.

2. Ellie is a juror on a criminal case. During the trial she is presented with evidence both in favor and against the defendant. Ellie is *uncertain* whether the defendant is guilty.

Both cases involve judgment under uncertainty, with a mixture of evidence supporting and opposing each event's likelihood. Yet, they involve what appears to be two qualitatively distinct representations of uncertainty. In the first case, Allie's uncertainty reflects the unpredictability inherent to a stochastic process (i.e., random draws from the pool of Bingo numbers). This type of uncertainty promotes a distributional mode of thought, with Allie perhaps thinking about how the outcome could play out in different ways upon similar occasions. In the second case, Ellie's uncertainty reflects the confidence she places in her beliefs, based upon what she knows about the details of the crime. This type of uncertainty promotes an evaluative mode of thought, with Ellie perhaps gauging the quality of the evidence, as well as what she knows and does not know. Allie reasons under what we will call *aleatory* uncertainty, while Ellie reasons under *epistemic* uncertainty.

The distinction between epistemic and aleatory uncertainty dates back to the early foundations of modern probability (Hacking, 1975). Probability theory is commonly thought to have originated in an exchange of letters between Blaise Pascal and Pierre de Fermat in 1654, over the question of how to properly divide the stakes in a game of chance were the game to be prematurely interrupted. To tackle this question, Pascal and Fermat formulated a calculus for how to think about events entailing aleatory uncertainty. Shortly thereafter Pascal posed the question of whether to believe in God as a decision-theoretic wager, with the outcome of the wager tied to the true state of nature (i.e., whether God in fact exists). In doing so, Pascal appropriated his earlier framework on the probability of chance events in order to understand a question entailing pure epistemic uncertainty. To this day probability theory is split between two dominant schools of thought, with Frequentists treating probabilities as the frequency of events repeated over multiple instantiations, and Bayesians treating probability as an index of subjective degrees of belief. Despite their differing interpretations, both rely on the same axiomatic foundation and therefore operate within the same set of constraints.

In this paper, we focus on how cognitive representations of epistemic and aleatory uncertainty affect the formulation and expression of quantitative judgment under uncertainty. Intuitively, it seems that epistemic and aleatory uncertainty differ in how they focus judgment. Epistemic uncertainty requires gauging one's confidence for events with a binary truth value (they are, or will be, either true or false). Aleatory uncertainty, on the other hand, entails evaluating propensities along a continuous unit interval. Because of this difference in focus, we hypothesize that judgments for epistemic events should be especially sensitive to differences in feelings of evidence strength. That is, whenever the balance of evidence favors one hypothesis over another we should expect greater judgment extremity under epistemic than aleatory uncertainty.

## Strength of Evidence and Judged Probability

In this section we use support theory (Tversky and Koehler, 1994) as a way to formalize the differences in judgment under epistemic and aleatory uncertainty. In support theory, probabilities are attached to *hypotheses*, or descriptions of events[1], with each hypothesis $A$ generating a non-negative support value, $s(A)$. Support values can be thought of as representing all feelings of evidence favoring a particular hypothesis — evoked by judgmental heuristics, extant knowledge, or anything else — much in the same way that utilities underlie a given preference ordering.

According to support theory, subjective probability is a function of the support generated for a focal hypothesis normalized relative to the support of its complement. That is, the judged probability that hypothesis $A$ holds rather than the complementary hypothesis $\bar{A}$, assuming one and only one obtains, is given by

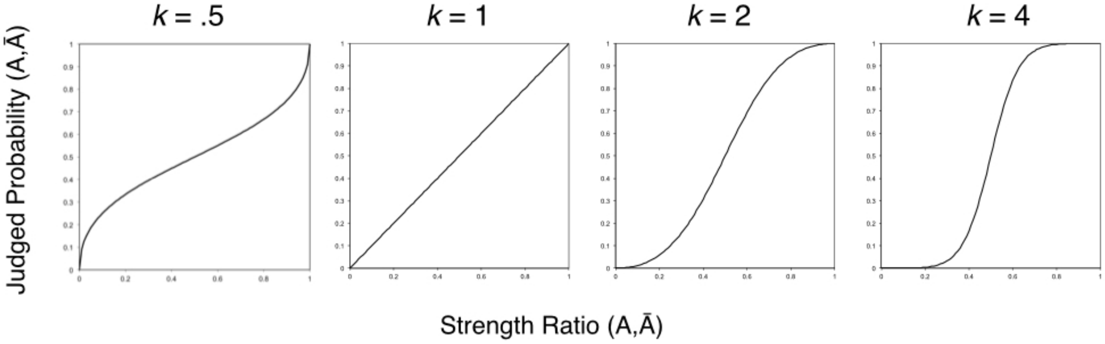
$$p(A, \bar{A}) = \frac{s(A)}{s(A) + s(\bar{A})} \tag{1}$$

Intuitively, one can think of probability judgment as determined by the balance of support for and against a particular hypothesis.

For our purposes, what is important about this model is how empirical assessments of evidence strength relate to latent support. Let $\hat{s}(A)$ be the empirically assessed strength of evidence favoring hypothesis $A$. We make two modest assumptions that have been empirically validated in prior research (Fox, 1999; Koehler, 1996; Rottenstreich and Tversky, 1997; Tversky and Koehler, 1994). First, we assume that direct assessments of evidential strength and support values (derived from judged probabilities) are monotonically increasing: $\hat{s}(A) > \hat{s}(\bar{A})$ iff $s(A) > s(\bar{A})$. In other words, hypotheses associated with relatively greater evidential strength should be viewed as more probable than hypotheses with less evidential strength. Second, corresponding strength and support ratios are monotonically related: $\hat{s}(A)/\hat{s}(\bar{A}) > \hat{s}(B)/\hat{s}(\bar{B})$ iff $s(A)/s(\bar{A}) > s(B)/s(\bar{B})$. That is, the higher

[1] 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).

Figure 1: Examples of Sensitivity to Evidence Strength ($k$)



the ratio of judged strength between the focal and alternative hypotheses, the higher the odds assigned to the focal hypotheses relative to the alternative hypothesis. If these two conditions hold, and support values are defined along a 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 $s(\mathrm{A}) = \hat{s}(A)^k$ (cf. Theorem 2 of Tversky and Koehler, 1994).

Intuitively we can think of the scaling constant $k$ as an index of an individuals' sensitivity to evidence strength, or how feelings of evidence are mapped onto a probability judgment. Put differently, even though judged probability should always increase whenever the balance of evidence favors that hypothesis, the *rate* at which it increases can vary — individuals can use differences in evidence strength to either make a bold or timid judgment. The scaling constant $k$ represents this rate of change. To illustrate this point, it is useful to first convert probabilities into odds. Using Eq. (1), assuming all probabilities are positive, and defining $R(A,\, \bar{A})$ as the odds that $A$ occurs rather than $\bar{A}$ (assuming only one obtains), we get

$$R(A, \bar{A}) = \frac{p(A, \bar{A})}{p(\bar{A}, A)} = \frac{\frac{s(A)}{s(A)+s(\bar{A})}}{\frac{s(\bar{A})}{s(A)+s(\bar{A})}} = \frac{s(A)}{s(\bar{A})} = \left[\frac{\hat{s}(A)}{\hat{s}(\bar{A})}\right]^k \tag{2}$$

As $k$ approaches 0, R($A$,$\bar{A}$) 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 $\hat{s}(A)/\hat{s}(\bar{A})$ and judged probability $p(A,\, \bar{A})$. As $k$ increases above 1 subjective probability will increasingly diverge to 0 or 1 as differences in evidence strength emerge (see Figure 1).

We expect that representations of uncertainty will influence sensitivity to evidence strength, with greater sensitivity when epistemic uncertainty is relatively salient. In other words, $k$ should be greater under epistemic uncertainty than aleatory uncertainty. To examine this, we take the natural logarithm of both sides of Eq. (2) to get

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

Using eq. (3) we can empirically estimate sensitivity to evidence strength by means of ordinary least squares regression, with the coefficient from the log strength ratio providing an estimate of $k$. In the studies that follow, we use this analysis strategy when probing for differences in sensitivity to evidence strength.

## Overview of Studies

All studies examine the hypothesis that participants provide relatively more extreme judgments, holding evidential strength constant, under epistemic uncertainty. In Studies 1a and 1b we compare judgments across a variety of domains that vary in their degree of "epistemicness" (i.e., relative amounts of epistemic and aleatory uncertainty). In Study 2 we compare judgments within domains, where participants provide estimates to forecasts that vary in epistemicness. In Study 3, we held all features of the task constant and experimentally induced feelings of epistemicness in subjects. Collectively, these studies examine whether perceptions of epistemic uncertainty lead to relatively more extreme judgments and greater sensitivity to evidence strength across a variety of tasks and settings.

## Study 1a

In Study 1a participants provided judgments across three domains — weather, professional basketball, and U.S. geography. We expected that questions about U.S. geography would be viewed as entailing primarily epistemic uncertainty, while predicting upcoming weather-related events would be viewed as entailing primarily aleatory uncertainty. We chose a third set of questions about future NBA basketball games that we anticipated would fall somewhere between the other two domains in its degree of epistemicness (especially among individuals with some degree of knowledge in the domain, such as basketball fans). Holding evidential strength constant, we expected probability judgment to be more extreme for domains high in epistemic uncertainty and less extreme for domains high in aleatory uncertainty.

### Study 1a Methods

The sample consisted of 37 participants[2] recruited from Amazon.com's Mechanical Turk (MTurk) who were self-identified as NBA fans. In return for completing an online survey participants given a small cash payment and entered into a drawing to receive an NBA basketball jersey of their choice. One participant was dropped from the analysis for using outside sources (e.g., Wikipedia) during the task. Subjects were on average 33 years old (range: 19–59 years), and 81% of the sample was male.

[2] We planned to sample 50 basketball fans, but were only able to obtain 37 participants before the NBA playoffs started. Note that the design of Study 1a was entirely within-subjects, so the sample still provided us reasonable statistical power.

Table 1: Study 1a Judgment Domains

| Domain | Sample 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? |

The study consisted of three phases. In the first phase participants answered 16 two-alternative questions on each of three topics: (i) outcomes of upcoming NBA playoff games, (ii) outcomes of upcoming temperature estimates for U.S. cities, and (iii) general knowledge questions about the geographic size of different U.S. states. For each question, one of the two alternatives was designated as the focal target[3], and participants were asked to estimate the likelihood (from 0% to 100%) that the focal target was more likely than the non-focal target to win their matchup (basketball), have a higher daily high temperature (weather), or was geographically larger (geography). Table 1 provides sample questions and Appendix A provides a complete list of targets. For each question-pairing, the choice of the focal target was counterbalanced across subjects. The ordering of judgment domains and questions within domains was randomized, with the only constraint that all judgments within a domain were to be completed before advancing to the next block.

The second phase of the study involved assigning strength ratings to targets. Following previous work (e.g., Tversky and Koehler, 1994), participants were provided with a list of the targets from each domain and were asked to scale them relative to the strongest target. For example, instructions for the basketball domain were as follows:

> Consider the eight basketball teams remaining in the NBA playoffs. First, choose the team you believe is the strongest of the eight, and set that team's strength to 100. Assign the remaining teams ratings in proportion to the strength of the strongest team. For example, if you believe that a given team is half as strong as the strongest team (the team you gave 100), give that team a strength rating of 50.

In the final phase of the study participants rated each domain for its degree of epistemicness. A single trial from each domain was selected at random and participants rated the question by indicating their level of agreement with 10 statements on 7-point scales (1 = *not at all*, 7 = *very much*). These items assessed both epistemic uncertainty ("determining the outcome to this question depends on knowledge or skill") and aleatory uncertainty ("the outcome to this question feels like it

[3] We chose this format for eliciting beliefs because it allows us to distinguish *overextremity* (the tendency to provide judgments that are too close to 0 and 100) from *overestimation* (the tendency to overestimate the likelihood of events). Traditional belief elicitation formats such as two-alternative forced-choice questions cannot distinguish between the two (see Brenner et al., 2005).

Table 2: Epistemic-Aleatory Rating Scale

1. The outcome to this question is in principle knowable in advance.
2. Determining the outcome to this question depends on knowledge or skill.
3. With enough information, one could know the answer to this question in advance.
4. The outcome of this question feels unpredictable. (R)
5. The outcome of this question has an element of randomness. (R)
6. The outcome of this question feels like it is determined by chance factors. (R)
7. The outcome of this question could play out in different ways on similar occasions. (R)
8. Well-informed people would agree on what the outcome to this question would be.
9. The outcome to this question has been determined in advance.
10. If I could consult an expert on this topic it would improve my prediction.

**Notes:** Participants rated each statement on 7-point scales (1 = *not at all*, 7 = *very much*). (R) = reverse coded.

is determined by chance factors"). Scale items were combined to form a single index of epistemicness, with higher items indicating that the task was viewed as entailing primarily epistemic uncertainty (Cronbach's $\alpha$ ranged from .83 to .88 across domains). Table 2 provides a list of all epistemicness items.

### Analysis Strategy

For all studies, probability estimates were recoded as decimals in the unit interval. For analyses that estimate sensitivity to evidence strength, probability judgments and strength ratings[4] were converted to a log odds metric, with judgments of complete certainty recoded as .001 and .999, respectively. As discussed in the introduction, transforming the data in this fashion allows us to derive estimates of sensitivity to evidence strength by means of OLS regression.

## Study 1a Results

Table 3 provides a summary of epistemicness ratings across the three domains. To our surprise, our NBA basketball fans rated basketball as the least epistemic domain, followed by temperature estimates, and then by geography questions. All means were reliable different from one another ($p$-values $< .01$). More importantly, if our hypothesis is correct then judgment extremity and evidence sensitivity ($k$) should follow a rank-ordering similar to epistemicness ratings — smallest for basketball, intermediate for city temperature, and highest for the state geography questions.

Indeed, we found that both judgment extremity and evidence sensitivity followed a similar pattern to ratings of epistemicness. To measure judgment extremity, we took the mean absolute deviation (MAD) from the ignorance prior ($p = .50$). Judgments were most regressive for basketball games (M = 0.19, SE = 0.01), middling for temperature estimates (M = 0.24, SE = 0.01), and most extreme for geography questions (M = 0.28, SE = 0.01). All means were reliably different

[4] In all studies we excluded a small number of trials where estimated probabilities fell outside of the 0-100 range, or where an item was given a strength rating of 0 (since this implies a misunderstanding of the ratio scale).

Table 3: Epistemicness Ratings and Judgment Extremity in Studies 1–3

| | | Judgment Extremity | | | | |
|---|---|---|---|---|---|---|
| | average<br>epistemicness | MAD from<br>$p = .50$ | median<br>$p > .5$ | median<br>$p < .5$ | proportion<br>$p = 0$ or 1 | proportion<br>$p = .50$ |
| *Study 1a* | | | | | | |
| Geography | 6.01 (1.09) | .28 | .90 | .10 | .33 | .20 |
| Temperature | 3.97 (1.21) | .24 | .80 | .20 | .15 | .21 |
| Basketball | 3.33 (1.12) | .19 | .70 | .30 | .10 | .18 |
| *Study 1b* | | | | | | |
| Geography | 6.09 (1.13) | .28 | .80 | .10 | .27 | .09 |
| Population | 5.90 (1.23) | .33 | .90 | .10 | .16 | .05 |
| Oceans | 5.74 (1.21) | .36 | .95 | .10 | .41 | .04 |
| Crime | 4.53 (1.55) | .31 | .80 | .20 | .13 | .07 |
| Housing | 4.04 (1.52) | .20 | .70 | .30 | .03 | .13 |
| Temperature | 3.19 (1.19) | .20 | .75 | .30 | .02 | .12 |
| Rain | 3.14 (1.23) | .15 | .62 | .30 | .01 | .16 |
| Movies | 3.09 (1.46) | .24 | .80 | .25 | .09 | .14 |
| Politics | 2.95 (1.13) | .16 | .60 | .35 | .05 | .13 |
| Baseball | 2.49 (1.28) | .12 | .67 | .30 | .02 | .41 |
| Football | 2.49 (1.05) | .10 | .65 | .30 | .00 | .49 |
| Soccer | 2.43 (1.17) | .11 | .65 | .40 | .02 | .37 |
| *Study 2* | | | | | | |
| Historic average | 5.02 (1.02) | .28 | .83 | .15 | .23 | .15 |
| Random day | 4.40 (1.02) | .25 | .80 | .20 | .10 | .16 |
| *Study 3* | | | | | | |
| Epistemic prime | 2.90 (1.47) | .20 | .75 | .25 | .05 | .21 |
| Aleatory prime | 2.80 (1.60) | .17 | .70 | .30 | .02 | .23 |

**Notes:** For epistemicness ratings, standard deviations are in parenthesis. MAD = mean absolute deviation.

Table 4: Estimates of Sensitivity to Evidence Strength in Study 1a

| | Model I | Model II | Model III |
|---|---|---|---|
| Geography | $3.36^a$ (0.18) | $4.82^a$ (0.68) | $6.84^a$ (0.31) |
| Temperature | $2.38^b$ (0.19) | $3.85^a$ (0.62) | $2.76^b$ (0.24) |
| Basketball | $1.71^c$ (0.16) | $2.34^b$ (0.34) | $1.50^c$ (0.42) |
| | | | |
| Unit of Analysis | trials | subjects | items |
| No. of observations | 1,716 | 108 | 96 |
| No. of groups | 36 | 36 | 52 |
| $R^2$ | .295 | .085 | .861 |

**Notes:** Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates ($p < .05$).

from one another ($p$-values $< .05$). Table 3 provides additional indices of judgment extremity (or conversely, judgmental timidity), including median probabilities above and below .50, the proportion of responses indicating complete certainty ($p = 0$ or 1), and the proportion of responses indicating complete uncertainty ($p = .50$). We see a similar pattern across all measures except for the proportion of completely uncertain responses.

Next we examined estimates of sensitivity to evidential strength. Table 4 presents the results, analyzed three different ways. Model I displays the average $k$ for each domain estimated from data using all measurement occasions (i.e., each subject-trial is used as a data point), with participants treated as a random-effect. Model II displays estimates at the subject-level by running separate regressions on each participant's set of responses per domain, and then taking the average of those estimates across participants. Model III displays estimates at the item-level by taking the median response for each possible item-pairing (i.e., the response profile for a theoretically representative subject), and then estimating $k$ for each domain. In all three analyses we found the expected pattern of results. Sensitivity to evidence strength was weakest for basketball predictions, intermediate for temperature estimates, and greatest for geography questions. For instance, in the item-level analysis (Model III) we see a roughly 4.6-fold increase in sensitivity to evidence strength when going from a domain relatively low in epistemic uncertainty (basketball) to a domain relatively high in epistemic uncertainty (geography).

Lastly, we examined the relationship between sensitivity to evidence strength and perceptions of epistemicness. According to our hypothesis, the two should be positively correlated. First, we examined this at the trial-level by regressing judgments onto strength ratios, epistemicness ratings, and the interaction between the two. As expected, sensitivity to evidence strength increased when the task was relatively high in epistemic uncertainty ($b_{intx} = 0.65$, SE = 0.06, $p < .001$). Based on the regression model, $k$ was estimated at 1.43 for judgments one standard unit below the mean in epistemicness, and 3.52 for judgments one standard unit above the mean, yielding a 2.46-fold increase in sensitivity to evidence strength. Second, we examined this relationship at the subject-level by examining the correlation between each subject's regression estimates and their epistemicness ratings, separated out by domain. We again found a positive relationship between the two ($r$ = 0.28, $p$ = .003). Finally, we examined this relationship at the item-level by taking the median judgment for each item and regressing it onto median strength ratios, median epsitemicness ratings, and the interaction between the two. Again, we found that sensitivity to evidence strength is more pronounced for items higher in epistemic uncertainty ($b_{intx} = 1.61$, SE = 0.14, $p < .001$).

## Study 1b

Study 1a provides initial evidence for the hypothesis that judgment is more extreme under epistemic uncertainty than under aleatory uncertainty. In Study 1b we sought to replicate and extend the effects to a wider array of domains.

### Study 1b Methods

The sample consisted of 206 participants recruited from MTurk. One participant reported using outside sources while completing the task, and was dropped from the analysis. Subjects were on average 33 years old (range: 18–80 years), and 56% of the sample was female.

The procedure was similar to that of Study 1A. Participants first provided judgments about six questions that were randomly sampled from a pool of 12 questions, with each question drawn from a unique topic domain (see Table 5). Next, they provided strength ratings for the two targets in each of their six probability estimates. Third, participants were presented with the same six questions they responded to earlier, and rated each question for its epistemicness using an abridged 4-item scale (sampled from the 10-item scale used in Study 1a).

### Study 1b Results

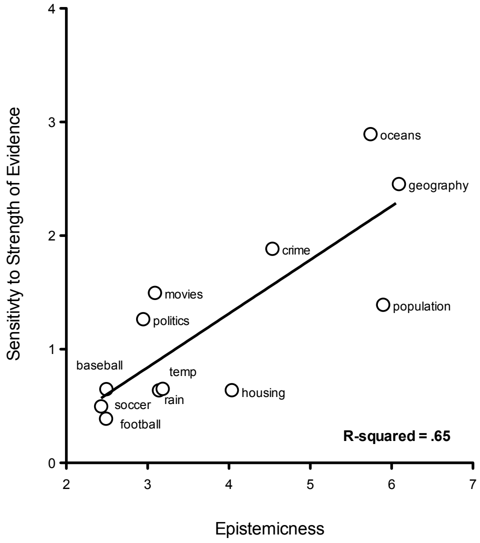
Table 3 lists the average epistemicness rating by domain, and also indices of judgment extremity. There was a wide range in epistemicness ratings, suggesting that perceptions of epistemic and aleatory uncertainty vary considerably from domain to domain. We also found that, as expected, participants were more likely to provide extreme probability judgments in highly epistemic domains. Using the mean absolute deviation from the ignorance prior, the correlation between epistemincess ratings and judgment extremity was positive and substantial ($r$ = .90). Average epistemicness ratings were also correlated with median judgments above and below .50, proportion of completely certain responses, and (inversely) to the proportion of completely uncertain responses (absolute correlations ranged from .73 to .93). At the individual-level, we find an average correlation of .42 between each subject's rank-ordering of epistemicness ratings and their rank-ordering of judgment extremity. All correlations were significant ($p$-values < .001).

Next, we estimated sensitivity to evidence strength for each of the 12 domains, with subjects treated as a random effect. Displayed in Figure 2, domain estimates of $k$ were highly correlated with the average epistemicness rating of each domain ($r$ = .80). That is, judgments were most sensitive to differences in evidence strength for domains entailing primarily epistemic uncertainty.

Table 5: Study 1b 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? |

Figure 2: Correlation between $k$ and Epistemicness in Study 1b



Overall, we see a 7.4-fold increase in $k$ when going from the domain lowest in epistemic uncertainty to the domain highest in epistemic uncertainty. We also examined sensitivity to evidence strength at the item-level by taking the median response, median strength ratio, and average epistemicness score for each of the 24 possible item-pairings. We then regressed probabilities onto strength ratings, epsitemicness ratings, and the interaction between the two. As expected, we find that the epistemicness associated with a particular question moderated the relationship between evidence strength and judged probability, $b_{intx} = 0.74$, SE $= 0.13$, $p < .001$. When a question was perceived to be relatively high in epistemic uncertainty, differences in evidence strength tended to lead to more extreme probability judgments. Based on these regression estimates, we would expect to see a 3.9-fold increase in evidence sensitivity when comparing questions one standard unit above and below the mean in epistemicness (2.70 and 0.69, respectively).

## Study 2

Studies 1a and 1b suggest that domains higher in epistemic uncertainty are associated with more extreme judgments and greater sensitivity to differences in evidence strength. One limitation of these studies is that different domains require different measures of evidential strength. It is unclear therefore, whether the (unobserved) measurement error associated with the elicitation of strength ratings is correlated with differences in sensitivity to evidence strength. It is possible that greater evidence sensitivity to evidence occurs in highly epistemic domains because these domains, for whatever reason, more readily lend themselves to assessing evidence strength along a ratio scale. Consequently, they provide a tighter fit to corresponding probability judgments, and therefore greater sensitivity to the evidence.

In Study 2 we hold the strength elicitation format constant while manipulating perceptions of epistemic and aleatory uncertainty. Participants compared U.S. cities along various weather-based attributes (temperates, rainfall, or smog levels), and estimated the likelihood that a designated city scored higher along a given attribute. Participants were asked to compare U.S. cities either based on historic averages or on a randomly selected day over the same time interval. In both formats participants provide likelihood estimates on the basis of the same strength attribute, so presumably the same feelings of evidence strength should come to bear on the judgment. Both formats, in other words, allow for a common strength elicitation procedure. The design of Study 2 therefore provides an "apples to apples" comparison on the utilization of evidence strength when arriving at a probability judgment. We expected that weather estimates for historical averages would be viewed as higher in epistemic uncertainty than weather estimates for randomly-selected days from the past. Accordingly, we hypothesized that participants would provide relatively more extreme judgments, and relatively greater sensitivity to evidence strength, in the historic average format.

The design of Study 2 also allowed us to test an alternative theoretical account for differences in judgment extremity. Brenner's (2003) random support model suggests[5] that overextremity arises from variability in support distributions. The logic is that feelings of evidence strength are randomly drawn from a normalized distribution of support, and greater variability in support distributions should result in more extreme probability judgments because highly discrepant support values will be more common. While our account does not contest these claims, it argues that judgment extremity can also arise due to differences in the mapping of support onto a probability judgment (i.e., sensitivity to evidence strength). By examining the variability in strength ratings (which serve as indirect proxies for support distributions) we can rule out this alternative explanation for any of the observed differences in judgment extremity.

## Study 2 Methods

A sample of 200 participants were recruited from MTurk. One participant was removed for reporting that they were less than 18 years of age. Subjects were on average 36 years old (range: 18–74 years), and 60% of the sample was female.

Subjects provided two-alternative probability estimates for U.S. cities in terms of relative rainfall, temperature, and smog. Each block consisted of 5 cities, with participants providing estimates for all pairwise comparisons within blocks (resulting in 10 estimates per block). Participants were asked to compare the pairs of cities either in terms of the historic average from the previous year (*historic average* format) or in terms of a randomly selected day from the previous year (*random day* format). To illustrate, sample questions from each condition were as follows:

> *Historic average*: "Consider the average amount of rainfall last year in Boston and in Anchorage. What is the probability that Boston had more rain last year?"

> *Random day*: "Consider a day picked at random last year in Boston and in Anchorage.

[5] It is worth noting that random support theory was designed to model calibration of subjective probability, and therefore most features of the model are irrelevant for the purposes of the current paper.

What is the probability that Boston had more rain that day?"

Similar to our previous studies, responses were made on a 0-100 scale, the ordering of blocks and questions within blocks was randomized, and the focal target for a given question was counterbalanced across participants.

Finally, participants then provided strength ratings of each city in terms of warmth, wetness, or amount of smog. In this study we did not probe for perceptions of epistemic and aleatory uncertainty, although a pilot sample using identical materials confirmed that historic average questions were consistently rated higher in epistemicness than random day questions[6]. Therefore, we expected that participants would provide more extreme probabilities, and more sensitivity to evidential strength, for the historic average format than for the random day format.

## Study 2 Results

As expected, participants gave more extreme judgments in the historic average format than in the random day format. The mean absolute deviation from the ignorance prior was greater when estimating historic averages (M = 0.28, SE = 0.01) than randomly selected days (M = 0.25, SE = 0.01), $t_{196} = 2.69$, $p = .008$. We also saw relatively greater judgment extremity for historic averages when comparing median probabilities above and below .50, and the proportion of completely certain responses ($p < .001$; see Table 3). As in Study 1a, we failed to find reliable differences in the proportion of completely uncertain responses ($p = .81$).

Next we examined for sensitivity to differences in evidence strength. Displayed in Table 6, sensitivity was more pronounced in the historic average condition than in the single-day condition. This is true when the results are analyzed at the trial-, subject-, or item-level. For instance, when aggregating across domains and comparing the results at the item-level (Model III), we see a roughly 1.4-fold increase in $k$ (when aggregating across domains) for participants who provided judgments to historical averages compared to participants who provided judgments for randomly-selected days from the prior year.

Lastly, we examined for differences in the use of strength ratings between the historic average and random day formats. If responding to historic averages (compared to random days) led to more variability in strength ratings, then this could potentially account for the observed differences in judgment extremity (consistent with a random support model; Brenner, 2003). We found no evidence, however, that question format affected evaluations of evidence strength. Using robust tests of variance with adjustments made for clustered data (Iachine et al., 2010; Levene, 1960), we failed to find reliable differences in the variability of focal strength ratings, foil strength ratings, or

[6] A sample of 157 participants (recruited from the same subject pool used in Study 2) received one question randomly drawn from each of the three domains, and rated that question using the same 10-item epistemicness scale used in Study 1a. Roughly half of participants viewed questions from the historic average condition, while the other half viewed questions from the single-day condition. As expected, participants viewed questions asking about historic averages as more epistemic (combined M = 5.02, SD = 1.02) than questions asking about a randomly-selected day from the previous year (combined M = 4.40, SD = 1.02), $t_{154} = 4.43$, $p < .001$. Perceptions of epistemicness reliably differed according to question format for all three domains ($p$-values $\leq .001$).

Table 6: Study 2 Results

| | Model I | Model II | Model III |
|---|---|---|---|
| *Rainfall* | | | |
| historic | 1.72$^a$ (0.06) | 1.50$^a$ (0.14) | 1.37$^a$ (0.08) |
| single-day | 1.29$^b$ (0.06) | 1.22$^a$ (0.15) | 1.45$^a$ (0.09) |
| *Temperature* | | | |
| historic | 3.35$^a$ (0.10) | 5.04$^a$ (0.42) | 6.02$^a$ (0.45) |
| single-day | 1.91$^b$ (0.10) | 2.36$^b$ (0.43) | 2.25$^b$ (0.45) |
| *Smog* | | | |
| historic | 2.09$^a$ (0.09) | 2.65$^a$ (0.23) | 2.38$^a$ (0.12) |
| single-day | 1.73$^b$ (0.09) | 1.95$^b$ (0.24) | 2.02$^b$ (0.11) |
| *All Domains* | | | |
| historic | 2.24$^a$ (0.05) | 3.06$^a$ (0.20) | 2.58$^a$ (0.22) |
| single-day | 1.56$^b$ (0.05) | 1.85$^b$ (0.21) | 1.79$^b$ (0.23) |
| Unit of Analysis | trials | subjects | items |
| No. of observations | 5,579 | 583 | 120 |
| No. of groups | 197 | 196 | 20 |
| $R^2$ | .375 | .037 | .633 |

**Notes:** Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates ($p < .05$).

overall strength ratios ($p$-values ranged from .241 to .903, and the observed $R^2$ was always less than .01). In short, the experimental manipulation did not have an appreciable impact on evaluations of evidence strength, but did influence how those feelings of evidence strength were mapped onto a probability estimate.

## Study 3

Study 2 demonstrated that questions higher in epistemic uncertainty lead to more extreme judgments and greater sensitivity to evidential strength, even when the elicitation procedure for assigning strength ratings was equated across conditions. In Study 3, we directly manipulate perceptions of epistemicness while holding all features of the judgment task constant, thereby providing a strong test of the hypothesis that perceptions of epistemicness causally influence the mapping of evidence strength onto judgment. To do so, we had participants perform a simple binary task where the underlying distribution is unknown. A common finding in this paradigm is that response patterns tend to match the underlying probability distribution (i.e., probability matching). Although this behavior is commonly-viewed as sub-optimal, recent research has suggested that probability matching may partly reflect an effort to discern underlying patterns in the data rather than to simply maximize payouts (Gaissmaier and Schooler, 2008; Goodnow, 1955; Unturbe and Corominas, 2007; Wolford et al., 2004). Accordingly, we varied the task instructions to either promote pattern detection (thereby making epistemic uncertainty salient) or to promote random guessing (thereby making aleatory uncertainty salient).

## Study 3 Methods

The sample consisted of 100 students recruited from a UCLA subject pool, who were each paid $5 for their participation. Participants were on average 20 years of age (range: 16–58 years), and 82% of the sample was female.

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 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 represented a fixed pattern (also see Gaissmaier and Schooler, 2008, for a similar design). The underlying proportion of events was the same in both blocks, with a 2:1 ratio for the more common event. The event designated as more common (Xs and Os), as well the order of the two blocks, was counterbalanced across participants. Lastly, participants received feedback about their prediction after each trial.

The key manipulation was how this first phase of the study was described to participants. In the *epistemic* 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 *aleatory* 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.

Thus, participants were prompted to think in ways that suggested looking for underlying patterns or random guesses (see the Appendix for the full set of instructions). To incentivize accuracy, participants were truthfully told that the most accurate participant would receive an additional $25 bonus.

In the second phase of the study participants provided 28 probability judgments to upcoming weather-related in 8 U.S. cities. For each trial, participants were presented with 2 cities (sampled from the pool of 8 cities), with one city designated as the focal city. Participants indicated the probability that the focal city would have the higher daily high temperature on the following

Table 7: Study 3 Results

| | Model I | Model II | Model III |
|---|---|---|---|
| Epistemic prime | $1.35^a$ (0.06) | $2.03^a$ (0.23) | $2.12^a$ (0.11) |
| Aleatory prime | $1.03^b$ (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 |

**Notes:** Standard errors in parenthesis. Column superscripts that differ indicate a statistically significant difference between estimates ($p < .05$).

July 1st. The order of the judgment trials was randomized, and the city designated as focal was counter-balanced across participants.

In the third phase of the study participants provided strength ratings (in terms of city "warmth") for the 8 cities, using the same procedure as before. In the fourth phase of the study, participants were presented with three of their trials from phase two, and rated each question on the 10-item epistemicness scale used in Study 1a. We averaged the three trials to form an index of perceptions of epistemicness for the judgment task (average Cronbach's $\alpha = .75$).

## Study 3 Results

As a manipulation check, we calculated average epistemic scores for each question and its complement. Questions were viewed as containing more epistemic uncertainty in the epistemic prime than in the aleatory prime (Ms = 4.16 vs. 4.03), but this difference was not statistically significant, $t_{99} = 1.23$, $p = .22$. As an internal analysis, we separately examined items measuring epistemic and aleatory uncertainty. First, we note that these two indices were weakly correlated with one another ($r = -.09$, $p = .37$). For the epistemic uncertainty subscale, we fail to see any differences between the two priming conditions (Ms = 4.69 vs. 4.80), $t_{99} = 0.90$, $p = .37$. However, we see a reliable difference for the aleatory subscale, such that participants viewed questions as higher in aleatory uncertainty for the aleatory prime than for the epistemic prime (Ms = 5.11 vs 4.63), $t_{99} = 2.82$, $p = .006$. Our manipulation, it appears, was more successful at priming aleatory uncertainty than epistemic uncertainty.

As expected, judgments were more extreme in the prediction task than in the guessing task. The mean absolute deviation from the ignorance prior was greater in the prediction task (M = .20, SE = 0.01) than in the guessing task (M = .17, SE = 0.01), $t_{98} = 1.76$, $p = .08$. Displayed in Table 3, median judgments above and below .50 were more extreme in the prediction task, and there was a greater proportion of completely certain responses ($p = .073$). Again, we did not find reliable differences between conditions in the proportion of completely uncertain responses ($p = .80$).

Also as predicted, we found greater sensitivity to evidential strength in the prediction task than in the guessing task. For all three analyses (trial-level, subject-level, and item-level), we find that sensitivity is greater under the epistemic prime than under the aleatory prime. For example, when comparing conditions at the item-level (Model III), we saw a 1.9-fold increase in sensitivity to evidence strength when participants were instructed to look for patterns than when they were instructed to "guess" from an underlying distribution.

The observed difference in sensitivity to evidence strength could not be accounted for by differences in the use of strength ratings. Using robust tests of variance with adjustments made for clustered data, no reliable differences emerged in the variability of focal strength ratings, foil strength ratings, or overall strength ratios ($p$-values ranged from .317 to .886, and the observed $R^2$ was always less than .01). As in Study 2, variability in strength ratings did not account for the observed differences in judgment extremity across conditions.

Lastly, we examined for the relationship between epistemicness and sensitivity to evidence strength. Because our manipulation check found only differences in perceptions of aleatory uncertainty, we use this subscale as our predictor variable of interest. (Using the full epistemicness scale yields similar but weaker results). We examined this at the trial-level by regressing judgments onto strength ratios, aleatory ratings (measured at the domain-level), and the interaction between the two. As expected, sensitivity to evidence strength increased when the task was relatively low in epistemic uncertainty ($b_{intx} = 0.65$, SE = 0.06, $p < .001$). Based on the regression model, $k$ was estimated at 0.90 for judgments one standard unit above the mean in aleatory uncertainty, and 1.35 for judgments one standard unit below the mean in aleatory uncertainty.

# General Discussion

We present evidence suggesting that judgment is more extreme under epistemic than aleatory uncertainty. This pattern was observed across different judgmental domains (Studies 1a and 1b), when feelings of evidence strength were matched across tasks (Studies 2 and 3), and when participants were "primed" to focus on epistemic or aleatory uncertainty (Study 3). These findings suggest that the uncertainty associated with a judgment task can affect the quantitative expression of that judgment. More broadly, this research suggests that lay intuitions about the nature of uncertainty may have downstream implications for judgment and choice.

In the remainder of this paper, we discuss theoretical implications and extensions.

## Calibration and Overconfidence

As mentioned in the introduction, a stylized fact in the forecasting literature is that overconfidence varies substantially across domains (Klayman et al., 1999; Ronis and Yates, 1987; Wright and Wisudha, 1982; Wright, 1982). For instance, Wright and Wisudha (1982) found that participants displayed more overconfidence when "postdicting" past-events than when predicting future events, even though accuracy was roughly equal in both tasks. Based on our studies, we believe that some of the variance in overconfidence may be due to differences in perceived epistemicness in

Table 8: Study 2 Overconfidence Results

| | Judged probability | Proportion correct | Over-confidence | Brier score | Brier Score Decomposition | | |
|---|---|---|---|---|---|---|---|
| | | | | | Reliability | Resolution | Outcome variance |
| *Rain* | | | | | | | |
| historic avg | .763 | .696 | .067 | .202 | .010 | .056 | .250 |
| random day | .738 | .690 | .048 | .058 | .048 | .006 | .016 |
| *Smog* | | | | | | | |
| historic avg | .758 | .572 | .186 | .285 | .048 | .011 | .249 |
| random day | .749 | .546 | .202 | .112 | .086 | .000 | .026 |
| *Temperature* | | | | | | | |
| historic avg | .823 | .820 | .003 | .119 | .003 | .133 | .250 |
| random day | .764 | .774 | −.010 | .076 | .007 | .061 | .132 |
| *All domains* | | | | | | | |
| historic avg | .781 | .695 | .087 | .203 | .011 | .056 | .250 |
| random day | .750 | .669 | .081 | .082 | .038 | .012 | .057 |

Table 9: Study 3 Overconfidence Results

| | Judged probability | Proportion correct | Over-confidence | Brier score | Brier Score Decomposition | | |
|---|---|---|---|---|---|---|---|
| | | | | | Reliability | Resolution | Outcome variance |
| Epistemic prime | .702 | .659 | .042 | .227 | .011 | .033 | .250 |
| Aleatory prime | .673 | .624 | .049 | .230 | .007 | .025 | .250 |

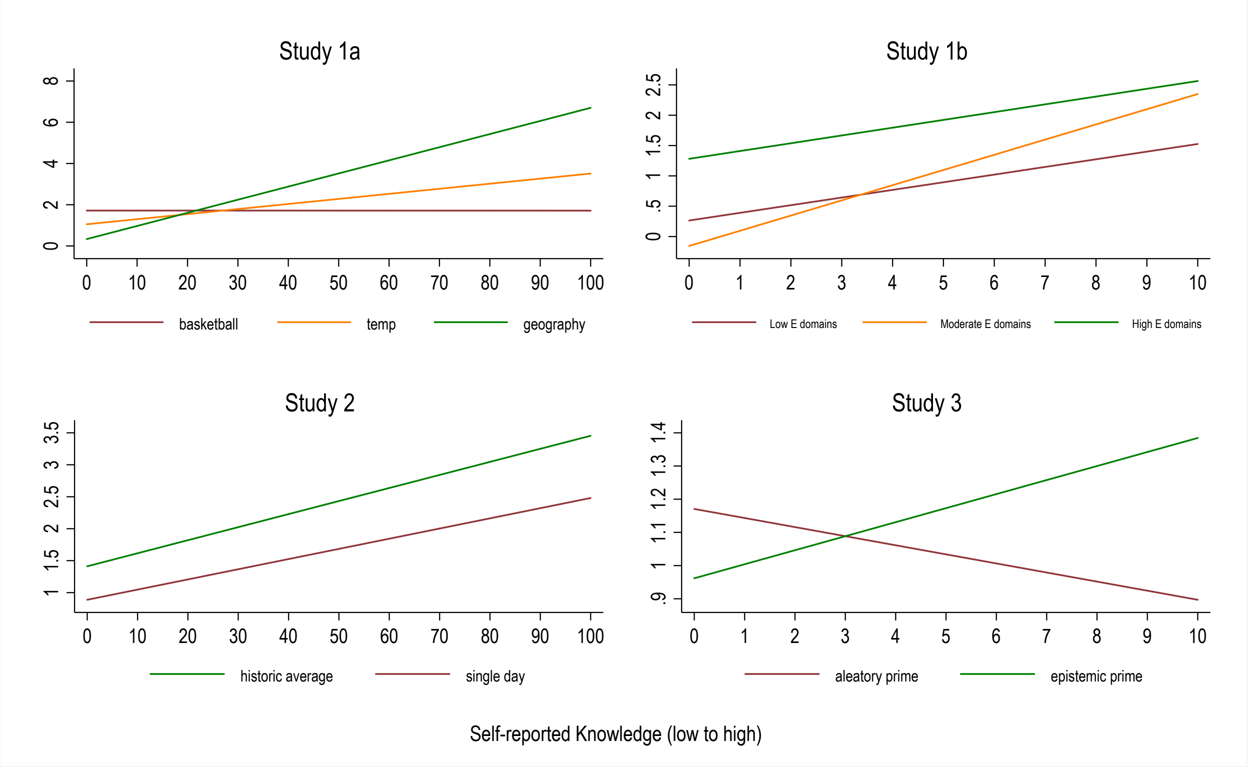
the task domain. If forecasters give more weight to the balance of evidence for highly epistemic domains, and participants are generally are overconfident, then on average they should be more overconfident in these domains. Accordingly, we re-analyzed the results from Studies 1b–3 to examine if overconfidence varied according to experimental conditions . . .

## Knowledge and Sensitivity to Evidence Strength

Our framework utilizes the original formulation of support theory. However, support theory does not directly accommodate the fact that people vary in degrees of knowledge or expertise. For example, people give more regressive probability estimates when they feel relatively ignorant about the task at hand (Fischhoff and Bruine De Bruin, 1999; Yates, 1982). Intuitively, it seems that feelings of subjective knowledge would also play a role in limiting the types of effects we report here. That is, if subjects feel completely ignorant about the judgment task, they are likely to give highly regressive judgments regardless of whether the task entails epistemic or aleatory uncertainty.

Future work could explore this possibility by using an ignorance prior model of judgment (Fox and Rottenstreich, 2003; Fox and Clemen, 2005), which incorporates the role of subjective knowledge alongside support in determining judgment under uncertainty. In the ignorance prior

Figure 3: Sensitivity to Evidence Strength as a function of Subjective Knowledge



model, probability judgments are represented by a weighted combination of support values and an anchoring on the ignorance prior (i.e., $1/n$ for $n$-alternative questions). More precisely, the judged odds $R(A,\,\bar{A})$ that hypothesis $A$ obtains rather than its complement $\bar{A}$ are given by

$$R(A,\bar{A}) = \left[\frac{n_A}{n_{\bar{A}}}\right]^{1-\lambda} + \left[\frac{\hat{s}(A)}{\hat{s}(\bar{A})}\right]^{k\lambda} \tag{4}$$

The second term on the right-hand side of eq. (4) represents the balance of support as measured by raw strength ratings, similar to eq. (2). The first term on the right-hand side represents the ratio of the ignorance prior for hypothesis $A$ and its complement, respectively. For two-alternative questions this returns odds of 1:1, for three-alternative questions this returns odds of 1:2, and so on. Finally, the parameter $\lambda$ represents the relative weight placed on the ignorance prior and support values, and takes a value between 0 and 1. As $\lambda$ approaches 1, increasing weight is placed on the balance of evidence (i.e., support values); as $\lambda$ approaches 0, judgments regress to the ignorance prior. Intuitively, one can think of $\lambda$ as indexing feelings of subjective knowledge. When people feel relatively ignorant, they are likely to anchor heavily on the ignorance prior. When people feel relatively confident in their knowledge, they tend to give little weight to the ignorance prior and instead rely on subjective impressions of evidence strength.

The ignorance prior model makes a clear prediction about the interaction between subjective knowledge and feelings of epistemicness: Any differences in sensitivity to evidence strength as a function of epistemicness should be greatest under high knowledge. When subjective knowledge is low, by contrast, all judgments should regress to the ignorance prior. For exploratory purposes, we included a single-item self-reported knowledge item for each judgmental task in all studies reported above. For each study, we examined the three-way interaction between evidence strength, task condition or domain[7], and self-reported knowledge on judgments. (Analysis was performed at the trial-level, with participants treated as a random effect). Figure 3 graphically displays the interaction derived from these regression estimates. In general, we find weak but suggestive evidence for the predicted interaction. For example, in Study 1a we see that sensitivity to evidence strength increases as a function of subjective knowledge most rapidly for geography questions, less rapidly for temperature forecasts, and least so for basketball matches. This is precisely the pattern we would expect to see given that geography, temperature, and basketball questions were respectively viewed as relatively high, medium, and low in epistemic uncertainty. These results, however, 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 statistically derived from multiple trails 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 (see See et al., 2006, for an example).

[7] For Study 1b, we reduced the 12 domains to three groups (low-, moderate-, and high-epistemic domains) by performing a tertiary-split on average epistemic ratings from each domain.

# 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.

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.

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.

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.

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.

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.

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.

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

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

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 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.