

Understanding others’ models of uncertainty

Joel M. Levin*

Craig R. Fox†

Gülden Ülkümen‡

Working paper, June 2026

Abstract

Prior work has established that people intuitively distinguish between uncertainty that is reducible given additional information (epistemic uncertainty) and uncertainty that is inherently irreducible (aleatory uncertainty) and that they use distinct language when describing each. In epistemic domains, people commonly use language such as “I am N% sure”, whereas in aleatory domains, language such as “There is an N% chance” is more common. We hypothesize that people use such language as a cue when evaluating others’ competence – preferring others who use language that is consistent with their own beliefs about the type of uncertainty that characterizes a given outcome. We find support for this pattern across two studies. In Study 1, participants rate others as being more competent when they use confidence *language* (such as “... N% sure”) when offering advice about outcomes that are perceived to be relatively more epistemic and *likelihood language* (such as “... N% chance”) when offering advice about outcomes that are perceived to be relatively more aleatory. In Study 2, we employ a manipulation to shift perceptions of uncertainty type within the context of a single, real-world outcome. Priming participants to think about an outcome as being relatively more epistemic or more aleatory shifts their predictions about the judgment accuracy of others who use confidence language versus those who use likelihood language.

Introduction

I’m thinking of a stock in the S&P 500. What is your best estimate about whether it will rise or fall tomorrow? Would you be any more confident in your estimate if you knew the name of the company? What if you could read the company’s financial reports? And what if you could confer with an expert?

That answers to questions like these vary across individuals is evidence of variation in underlying beliefs about the *type* of uncertainty surrounding stock movements. Answering “yes” to one or more

*Rady School of Management, University of California San Diego. Email: joelmlevin@gmail.com.

†Anderson School of Management, University of California Los Angeles.

‡Marshall School of Business, University of Southern California.

of the questions reveals that a person has a model of uncertainty that is at least somewhat *epistemic*. Epistemic uncertainty arises from salient missing information, such as the name of a company, its financial performance, or features that you might learn by talking to an expert. Because epistemic uncertainty arises from missing information, it can be reduced, in full or in part, by obtaining some of that information. Likewise, the more a person answers “no” to questions of this type, the more they reveal that they have a relatively *aleatory* model of uncertainty. Aleatory uncertainty arises from factors like stochasticity, complexity, and unobservability, and thus cannot be reduced by obtaining additional information¹.



More epistemic	More aleatory
arising from a salient missing knowledge or information	arising from stochasticity, complexity, unobservability, (etc.)
reducible by obtaining that knowledge/information	inherently irreducible

Figure 1: Study 1 and Study 2 outcome question stimuli overview

These fundamental differences in beliefs influence our judgments in powerful ways. The extent to which we perceive uncertainty as being relatively more epistemic or aleatory affects how we weigh evidence (Tannenbaum, Fox, & Ülkümen, 2017), the kind of information that we search for (Walters, Ülkümen, Erner, Tannenbaum, & Fox, 2020), our risk preferences (Fox, Goedde-Menke, & Tannenbaum, 2021), how we judge others’ expressions of uncertainty (Levin, 2026), and how we attribute credit and blame (Fox, Ülkümen, Tannenbaum, Walters, & Erner, 2022).

Moreover, holding an inaccurate model of the kind of uncertainty that you’re facing can be consequential. The present domain offers a clear example: consider the decision of whether to invest your retirement savings in actively or passively managed funds. People who hold relatively more

¹Base rates are a special category of information.

epistemic models of the uncertainty surrounding the performance of individual stocks are often willing to pay fees in exchange for having their investment portfolios managed by professionals who pick winners and losers among individual stocks. In contrast, people who hold relatively more aleatory models of uncertainty – perceiving the *relative* performance of individual stocks to be effectively-random – more often prefer investment strategies that forgo active management and are therefore able to maintain lower costs (Walters et al., 2020). In this domain, making the (historically) wrong decision can mean that a person needs to work for several additional years before being able to afford retirement (Carhart, 1997).

To illustrate the costs of holding inaccurate models more generally, consider two contrasting errors. First, suppose that a hypothetical outcome is fundamentally aleatory – driven by inherent randomness and complexity – but that you believe it to be epistemic. In such a case, your belief that the outcome is relatively epistemic would cause you to focus on characteristics of the present situation – searching for patterns and predictive signals where none meaningfully exist. In an extreme case, a person holding an epistemic model of an aleatory outcome might spend hours analyzing the specific circumstances related to the week’s lottery numbers, convinced that careful study will enable them to predict the outcome of the lottery at rates better than chance.

On the other hand, consider the (less common) case in which a person believes that an outcome is more aleatory than it is in reality. Here, inaccurate beliefs about uncertainty type would cause one to overweight base rates and historical distributions while failing to account for potentially diagnostic information about the present case. For example, a person might perceive changes in their health as being inherently aleatory, dismissing the coming and going of an intermittent cough without attending to how its timing coincides with exposure to pollutants, like dust or smoke.

Incorrect beliefs about uncertainty type models can also directly distort probability judgments themselves – causing us to actually become more or less certain about the probability of an outcome, with implications for calibration and decision quality. Tannenbaum and colleagues (2017) document this phenomenon in the context of basketball predictions. When fans were asked to predict the winner of an upcoming game that was framed as a single event, they provided more extreme probability estimates than when the same game was framed as one contest in a longer series – a framing that was designed to naturally cue distributional thinking (which is inherently more aleatory

in nature.)

Given the importance of holding accurate models of uncertainty type for decision quality, decision makers would be well-served to seek information and advice from others who themselves hold accurate models. But how can we identify such others? Fortunately, recent research suggests that people's beliefs about uncertainty type "leak out" through the language they use when expressing probabilistic judgments, which may allow others to make inferences about those beliefs without them being described explicitly. Ülkümen, Fox, & Malle (2016) found that when people hold epistemic models of uncertainty, they tend to convey probabilistic claims using *confidence language* such as "I am 80% sure." When they possess aleatory models, they more commonly use *likelihood language* such as "There is an 80% chance."

Drawing on prior work, we assert that these linguistic cues reflect deeper cognitive differences in how people conceptualize uncertainty. Confidence language may signal that the speaker believes a predictive model exists and therefore that their uncertainty stems largely from missing information that could be used to populate that model. Likelihood language, by contrast, may indicate that the speaker views the outcome as fundamentally governed by chance processes that cannot be meaningfully predicted by gathering additional information.

The present research investigates whether people are sensitive to these linguistic cues when evaluating others and whether they use them to make consequential judgments about others' competence and the quality of their advice. If people intuitively prefer advisors whose language aligns with their own beliefs about uncertainty type, this would suggest a sophisticated understanding of the relationship between language, mental models, and decision quality.

Study 1

In Study 1, we sought to investigate how the type of language that people use to communicate uncertainty when conveying advice affects how others perceive them.

The preregistration can be found at <http://aspredicted.org/blind.php?x=r33zq7>.

Participants

A total of 237 participants (51% female; age: $M = 20.06$ years, $SD = 1.48$) were recruited from a university behavioral lab to participate in exchange for course credit. Students were recruited over a period of one week and the authors were not involved in the recruitment process. We preregistered a single exclusion criterion intended to screen out inattentive participants.² This criterion excluded a total of 7 participants, yielding a final sample of 230.

Procedure

Participants answered a series of questions and received hypothetical advice to aid them with each. After answering each question, participants rated the quality of advice and its source and indicated their beliefs about the type of uncertainty that characterized the question. The study employed a 9 (questions) x 2 (language: confidence / likelihood) mixed design. All participants answered all nine questions (in random order) and were randomized within each question to receive advice that contained either likelihood or confidence language in conveying probabilistic advice.

Question manipulation: We developed nine questions that were intended to be accessible to undergraduate students and that represented a wide range of perceived epistemicness. These questions are presented in Table 1, in descending order of epistemicness as rated by participants in this study.

Language manipulation: The advice that participants received for each question was conveyed in the form of a binary estimate or prediction accompanied by a numerical probability, following recent work on perceptions of uncertain advice (Gaertig & Simmons, 2018; Levin, Gilbertson, & Broomell, 2019). An example piece of advice is: “I am N% sure that the height of the empire state building is greater than 1000 feet” (*emp* question, confidence language stem). We elected to quantify the amount of uncertainty in each statement using numerical probabilities rather than verbal expressions of probability (e.g., “I am very sure that. . .”) because doing so helped us to hold

²We generated a participant-level variable by calculating the difference between rated epistemicness for the two questions that were rated by the sample as being most epistemic (*emp*, *pop*: mean epistemicness: 6.48) and for the two questions that were rated by the sample as being least epistemic (*die*, *steelers*: mean epistemicness: -4.05; mean difference = 10.53). We then calculated the standard deviation of this difference (3.13) and excluded participants whose differences were more than two standard deviations below the sample mean (mean difference < 4.27; the 3rd percentile).

the *amount* of uncertainty expressed by each advisor as constant as possible across language stems. For each question, participants received advice that contained one of six language stems: three containing confidence language and three containing likelihood language. The three confidence language stems were: “I am N% sure”, “I am N% confident”, and “I can say with N% certainty” and the three likelihood stems were: “It is N% likely”, “I’d say that there is an N% probability”, and “I think there is an N% chance”. The remainder of each piece of advice, including the numerical probability, was held constant across conditions. As preregistered, we collapsed the stems to form two conditions.

Table 1: Questions used in Study 1

Question (<i>abbreviation</i>)	Answers	Epistemicness (sd)
Is the height of the empire state building greater than or less than 1000 feet? (<i>emp</i>)	Greater than 1000 feet / Less than 1000 feet	6.87 (1.93)
Which country has a larger population? (<i>pop</i>)	Brazil / Russia	6.09 (2.33)
How long would it take for a brick dropped from the top of the Cathedral of Learning to hit the ground? (<i>cath</i>)	Greater than 3 seconds / Less than 3 seconds	4.38 (3.35)
What percentage of incoming 2019 CBA (college of business administration) students will score at least 1400 on the SAT? (<i>sat</i>)	More than 25% / Less than 25%	1.28 (3.32)
Will Donald Trump be reelected in 2020? (<i>trump</i>)	Yes / No	-0.55 (3.36)
If we randomly select another student in this room, will they be born in the same year as you? (<i>year</i>)	Yes / No	-1.36 (3.37)
Will it rain in Pittsburgh on July 4th, 2019? (<i>rain</i>)	Yes / No	-2.90 (3.30)
Will the outcome of rolling a six-sided die be greater than 2? (<i>die</i>)	Yes / No	-3.27 (3.55)
Will the Pittsburgh Steelers make the playoffs in the 2039 season? (<i>steelers</i>)	Yes / No	-4.83 (2.92)

On questions for which there was an objectively correct *probabilistic* answer (*year*, *die*), the advisor always suggested the correct answer and communicated a numerical probability that was objectively calibrated. On questions for which there was an objectively correct *absolute* answer (*emp*, *pop*, *cath*), the advisor always suggested the correct answer and the numerical probability contained in the advice was calibrated based on a pretest of other undergraduate students at the same institution. On questions about future events, for which there was no objectively or probabilistically correct

answer (*sat, trump, rain, steelers*), both the answer and the probability contained in the advice were calibrated based on a pretest of undergraduate students.

Measures

Advice / advisor quality: Participants rated the quality of each advisor and their advice by responding to the following three items: “How competent is this advisor?”; “How much do you trust this advisor?”; “Please rate the quality of this advice.”. The first two items used 7-item Likert response scales bounded by *not at all* and *extremely* and the third item used a 7-item Likert response scale bounded by *lowest quality* and *highest quality*. These items were selected because they were the three highest-loading items from the 7-item composite measure used in Gaertig & Simmons (2018) and were found to have high reliability in subsequent work (Levin et al., 2019). The dependent measure is a simple average of the three items.

Epistemicness: Participants reported their beliefs about the epistemicness of each question using the four-item version of the Epistemic-Aleatory Rating Scale (EARS: Fox, Tannenbaum, & Ülkümen, 2016). The EARS measures beliefs in two dimensions of uncertainty: epistemic (uncertainty due to factors that are in-principle knowable) and aleatory (uncertainty due to factors that are random in nature). The two items for epistemic uncertainty were (the belief that the question. . .) “is knowable in advance, given enough information”, and “is something that well-informed people would agree upon”. The two items for aleatory uncertainty were (the belief that the question. . .) “is something that has an element of randomness”, and “is determined by chance factors”. To generate an index of epistemicness, we reverse-coded the aleatory items, as in Tannenbaum et al. (2017). All items used 5-item Likert response scales bounded by *not at all* and *very much*. The scale has theoretical bounds of [-8, 8], with larger (positive) values indicating the belief that a question is more epistemic and smaller (negative) values indicating the belief that a question is more aleatory.

Results

We hypothesized that participants would rate advice containing confidence language more favorably (compared to advice containing likelihood language) when they perceived questions as being more epistemic and that participants would rate advice containing likelihood language more favorably

(compared to advice containing confidence language) when they perceived questions as being less epistemic. Our primary analysis, which we specified in advance, was a regression of the specification: $quality_{iq} = \alpha + \beta_1 epistemicness_{iq} + \beta_2 language\ type_{iq} + \beta_3 epistemicness_{iq} * language\ type_{iq} + \epsilon_i$, with robust standard errors clustered at the participant level. We predicted a significant interaction between language type (β_2) and perceived epistemicness (β_1), which would indicate that the relationship between the type of language that a person uses when giving uncertain advice and the degree to which they are perceived favorably depends on their target's beliefs about the epistemicness of the question at hand.

Advice / advisor quality:

This section proceeds as follows: First, we report main effects of *question*, *language type*, and *perceived epistemicness* on *advice / advisor quality*. Then, we report and discuss the interaction between language type and perceived epistemicness, which constitutes our primary test. Finally, we report the effects of language type on quality at the question level. Throughout this section we refer to advice / advisor quality simply as *quality*.

We observe significant heterogeneity in quality ratings across questions, $F(8) = 33.07$, $p < .001$. The question for which participants rated advisors and advice most favorably had a mean quality rating of 4.12, 95% C.I. = [3.94, 4.29] while the question for which participants rated advisors and advice least favorably had a mean quality rating of 2.42, 95% C.I. = [2.26, 2.59]. We attempt to empirically account for this heterogeneity by conducting all rating-level analyses both with and without question fixed effects.

We find a directionally positive but insignificant main effect of the use of language type on quality, $\beta = .10$, $t(2068) = 1.38$, $p = 0.17$, indicating a weak preference for likelihood advice across the questions in our study. However, this relationship depends on the nature of questions asked and is therefore not interpretable without regard for them. Of greater interest is within-question variation, which is visualized in *Figure 3* and addressed in subsequent discussion. The main effect of language type on quality is similar (and similarly uninterpretable) when including question fixed effects, $\beta = .10$, $t(2060) = 1.54$, $p = 0.12$.

We observe a positive and significant main effect of perceived epistemicness on quality, $\beta = 0.71$,

$t(2068) = 10.92$, $p < 0.001$. The strength of this relationship attenuates but remains highly significant when controlling for question characteristics using question fixed effects, $\beta = 0.65$, $t(2060) = 5.46$, $p < 0.001$. There are two interesting things to note about this pattern. First, the inclusion of question fixed effects accounts for the different numerical probabilities contained in each piece of advice (which is constant within-question for all participants), which rules out potential explanations based on perceived fit between stated probabilities contained in advice and perceived epistemicness (Levin et al., 2019). Second, that the relationship between perceived epistemicness and quality is not eliminated when introducing question fixed effects suggests that the heterogeneity in quality ratings across questions is not explained by differences in perceived epistemicness.

Consistent with our hypothesized pattern, we observe a significant interaction between perceived epistemicness and language type on quality, $\beta = -0.03$, $t(2066) = 2.39$, $p < 0.02$. Because the heterogeneity in quality ratings across questions does not appear to be associated with perceived epistemicness at the individual level (and is therefore a source of measurement error) the introduction of question fixed effects does not affect the coefficient of the interaction and modestly reduces its standard error, $\beta = -0.03$, $t(2058) = 2.64$, $p < 0.01$. This pattern holds when looking only at the effects of the two aleatoryness items ($\beta = -0.10$, $t(2058) = 2.20$, $p < 0.03$) or the two epistemicness items alone ($\beta = -0.12$, $t(2058) = 2.33$, $p < 0.02$), which suggests that participants who perceived a given question as being either more aleatory or less epistemic assigned higher quality ratings to advice that contained likelihood language (as opposed to confidence language).

To visualize this pattern in the presence of question-level heterogeneity, we generate a question-adjusted quality measure from each participant's raw quality rating by subtracting the mean quality rating from each, yielding a number that represents the degree to which a participant rated the advice higher or lower than other participants' ratings for the same question. This pattern is illustrated in *Figure 2*. We find that when participants receive advice that contains likelihood language, they rate advice and its source similarly regardless of their beliefs about the type of uncertainty that characterizes the question. However, when participants receive advice that contains confidence language, they rate advice much more favorably when they perceive the questions as being more epistemic and much less favorably when they perceive a question as being more aleatory.

Finally, we turn to patterns in preference for confidence versus likelihood language within each

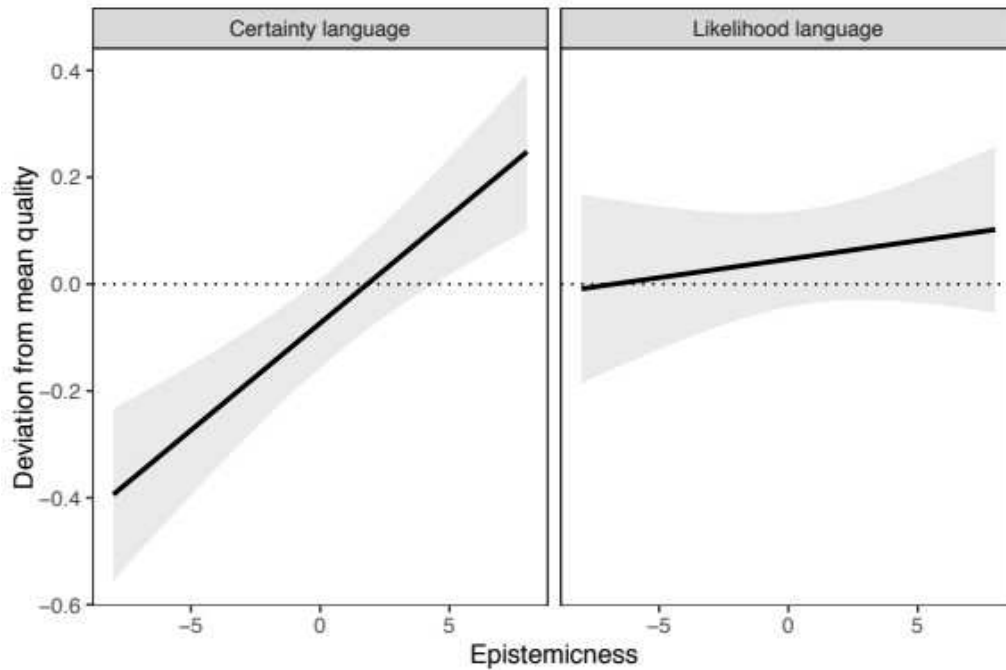


Figure 2: Study 1: fitted observation-level results from a linear model. The y axis represents question-adjusted quality, which is generated by subtracting mean question-level quality ratings from participants' individual quality ratings for each question. Shaded areas represent 95% confidence intervals.

of the questions in our paradigm. *Figure 3* presents both of the variables that are relevant to our hypothesis. On the right axis and denoted by grey squares are mean epistemicness ratings for each question, which are presented in descending order from left to right. On the left axis and denoted by black circles are differences between rated quality in confidence language versus likelihood language conditions, which is generated by subtracting the latter from the former at the question level. Thus, positive values indicate that participants rated advice and advisors for a given question more favorably when the advice contained confidence language and negative values indicate that participants rated advice and advisors for a given question more favorably when the advice contained likelihood language. The correlation between mean epistemicness and preference for confidence language is substantial, $r=.88$, $t(7) = 4.99$, $p < .002$.

Discussion

Study 1 offers support for the hypothesis that observers evaluate sources in part on a basis of the fit between the type of language that those sources use and their own beliefs about the type of

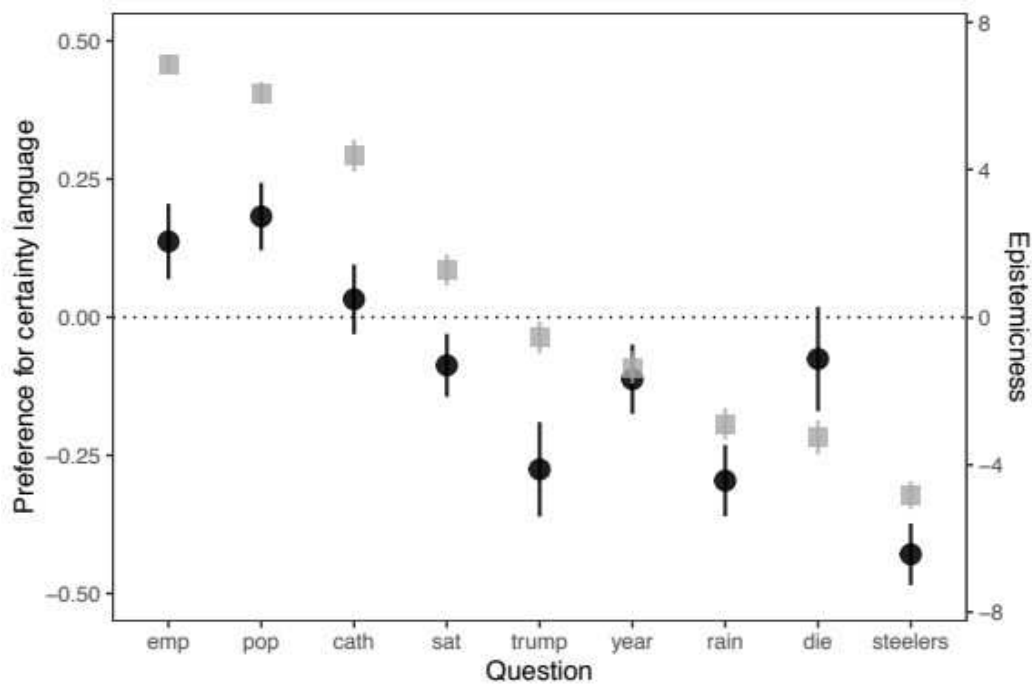


Figure 3: Study 1: question-level results. Grey squares denote rated epistemicness (y axis, right). Black circles denote the mean difference between rated quality for confidence versus likelihood language conditions, with positive values indicating higher quality ratings for confidence advice and negative values indicating higher quality ratings for likelihood advice (y axis, left). For both, vertical lines denote 95% confidence intervals.

uncertainty that characterizes the environment.

This study is limited in two notable ways. First, we make our key inferences on the basis of a comparison between observers' *endogenous* beliefs about uncertainty type and the type of language that sources use when describing uncertainty. This is a limitation because observers' beliefs about uncertainty type may be correlated with other endogenous characteristics that affect their perceptions of sources. Study 1 is also limited in that it relies on survey-based dependent measures. Although these measures allow us to capture relevant beliefs, these beliefs may or may not translate into consequential behaviors. Study 2 is designed to address both of these limitations.

Study 2

Study 2 took place on the campus of Harvard University in the spring of 2020. Each spring, first year students at Harvard College enter a lottery to determine in which of twelve Houses they will live for the remainder of their undergraduate careers.³ The Rising Sophomore Lottery (hereafter: *RSL*) was originally instituted in 1995 and serves to ensure that each House contains a diverse group of students that is representative of the College as a whole. As implied by the name, the RSL is random: the official Harvard RSL webpage clearly states that the RSL is “administered randomly using a basic mathematical process” and that “the only control in the lottery is for gender balance in each House” (Dean of Students Office, 2020).

Despite this description, many of the Harvard College students who we interviewed described their beliefs about the uncertainty surrounding outcomes in the RSL in a manner that reflected more ambiguous beliefs about the nature of uncertainty surrounding outcomes of the RSL. These beliefs are well documented in the public domain. In fact, at project inception, the first Google result for “harvard housing lottery” was not the official University page but rather a student-published essay titled “Why The Harvard Housing Lottery Might Not Be Random” (DiCara, 2020). Such uncertainty about whether the RSL is or is not random was what drew us towards conducting an experiment in this context. To the extent that the participants in our sample held ambiguous or conflicted beliefs about the RSL's randomness, we anticipated that we might be able to shift those

³There are two types of Houses: River Houses and Quad Houses. River Houses are better located and are widely considered to be more desirable for this reason.

beliefs via a framing manipulation.

Understanding the nature of uncertainty surrounding RSL outcomes is important for students because they face a number of decisions in preparing their application materials. First, each student has the option of selecting up to seven other students to form a *blocking group*. Students who are part of the same blocking group are guaranteed to be placed in the same House and will generally be placed in shared or adjoining rooms. Additionally, each blocking group may elect to *link* to up to one other blocking group, and linked groups are guaranteed to be placed in neighboring Houses. Finally, some students believe that characteristics of blocking groups' composition can affect RSL outcomes. Specifically, students told us that groups with larger numbers of varsity athletes or legacy students might receive preferential treatment when it comes to being assigned to River Houses. We found similar sentiments in public, online discussion groups.

The details of the design for Study 2 are relatively complex, but they belie a simple structure. We asked one group of Harvard College undergraduate students to make probabilistic predictions about a series of RSL outcomes using natural language. From these predictions, we manually selected two sets to serve as stimuli: one containing likelihood language and one containing confidence language. We then showed pairs of these predictions to another group of Harvard College undergraduates and asked them to make incentivized judgments about the overall accuracy of those students to whom the predictions belonged. Prior to this step, we assigned participants to one of two conditions in which they were primed to think about the RSL as being more or less random. A detailed description follows. The preregistration for Study 2 can be found at <https://aspredicted.org/blind.php?x=ut75wk>. Data, materials, and code are available upon request and will be posted publicly prior to publication. All hypotheses and statistical tests pertain only to the focal sample (described below).

Participants

Study 2 involved three unique and non-overlapping samples. The first sample, *the applicants* ($N = 7$), were a group of first year students who shared their RSL applications with us and consented to be contacted again after the resolution of the RSL to share their lottery outcome. The second sample, *the predictors* ($N = 71$), made probabilistic predictions about RSL outcomes for all 7 of the *applicants* which they conveyed using natural language. The third and final sample, *the focal sample*

($N = 525$; 62% female, 1% non-binary; mean age = 19.9 years, $SD = 1.83$), made incentivized predictions about which of the *predictors* they thought would be most accurate. All participants were recruited by the Harvard Decision Science Lab.

Procedure

The applicants

Each participant gave us the details of their RSL application in a standardized format. After submitting their unique Harvard University ID and consenting to be contacted again after the resolution of the RSL, participants told us: whether they were in a blocking group, the size of their blocking group, whether their blocking group was linked to another blocking group, as well as the number of women, varsity athletes, and legacy students in their blocking group.

The predictors

The purpose of the predictor sample was to generate stimuli to be used in the focal portion of the study. *Predictors* were recruited by the Harvard Decision Science Lab to complete a study called “Predict who gets Quadded!⁴”. We asked *predictors* to make probabilistic predictions about whether each of the seven *applicants* would end up in a River House or a Quad House. We scored participants using a scoring system based on inverse Brier scores, which we explained carefully and communicated visually using a graph. In practice, participants earned more points to the extent to which their judgments were accurate and well-calibrated. Participants were paid cash bonuses based on their performance.

We required participants to make each of their predictions in the form of a complete sentence.⁵ To maximize our chances of obtaining a sufficiently diverse set of stimuli, we randomized one element of the instructions such that participants either saw example sentences that contained likelihood language or example sentences that contained confidence language. Most participants made predictions that contained a probability of being placed in a River House. For this reason, we

⁴To get “Quadded” is slang for being assigned to a Quad House rather than a River House.

⁵Participants were free to enter anything into the open ended text boxes which we used to elicit their predictions, but were informed that they would only earn a bonus if they conformed to the rules by expressing their prediction as a complete sentence.

selected only from such responses. The stimuli which we selected are presented in Table 2. These stimuli were selected manually by the authors and they are not intended to be representative of all of the responses in our data.

Of note, we aimed to select stimuli that were relatively well balanced on those dimensions of uncertainty which we are not intending to study directly.

Table 2: Study 2 language stimuli

Confidence Language Predictions	Likelihood Language Predictions
“I am __% certain they will end up in a River House.”	“I would say there is a __% chance this student ends up in the River House.”
“I am __% sure that this student will be in a River House.”	“I would give them a __% probability of ending up on the River.”
“I am __% confident that (initials) will get a River House.”	“I think there is a __% likelihood that (initials) gets a River House.”

The focal sample

Participants in the focal sample were recruited by the Harvard Decision Science Lab to complete a study called “Predict who can predict who gets Quadded!”. We first described to participants the first two waves of data collection, including the nature of quadratic scoring rules and how *predictors*’ predictions were scored. We then informed them that their task was to “compare pairs of predictors and to estimate which of them will score the most points.”⁶ In each of three rounds, we showed participants one pair of predictions, each made by a different predictor, and asked them to estimate which of the two predictors was more accurate. In all rounds, one prediction contained likelihood language and the other prediction contained confidence language. We counterbalanced all stimuli such that there was an equal probability that participants would see any possible pair of contrasting statements and we further randomized which statement appeared on the right or left of the screen.

All prediction stimuli were presented with the numerical probability redacted, as presented in Table 2. Although this makes the predictions somewhat less realistic, we decided that the marginal decrease in realism was warranted given that this allows us to address some threats to internal validity

⁶Where points were scored as an inverse function of Brier score. This was described to participants.

which were present in Study 1. To make this feature less surprising, we informed participants that numerical percentages would be redacted in advance of their first prediction.

In addition to the within-subjects randomizations, Study 2 employed a single between-subjects manipulation. Following the instructions but prior to the first prediction, we asked participants one or two questions that were designed to cue either epistemic or aleatory thinking. For participants in the epistemic condition, we designed a manipulation that was intended to increase the salience of individual students' attributes and the role that those attributes could play in reducing uncertainty around the outcome of the RSL.⁷ To accomplish this, we showed participants the profiles of two hypothetical RSL applicants and asked them which they thought would end up in a River House.⁸ Conditioned on their choice, we asked them to write "one thing that (the other student) could do to potentially improve their ability to get into a River House." We required that participants enter at least one character into the text box in order to continue but did not require or encourage any additional effort. This stimulus is shown in Figure 4.

For participants in the aleatory condition, we designed a manipulation that was intended to cue distributional thinking and increase the salience of base rates, both of which are hallmarks of relatively more aleatory models of uncertainty. To do this, we simply asked participants to answer the question: "If we picked 100 first year students, how many of them would you expect to end up in River Houses?" As in the epistemic manipulation, we elicited participants' response using an open ended text box. As such, Study 2 employed a 2 (framing: epistemic / aleatory) x 2 (language: confidence / likelihood) mixed design, with framing manipulated between subjects prior to predictions and language manipulated within subjects in each of three rounds.

Measures

Epistemicness

We measured perceptions of the uncertainty surrounding the RSL using the six-item EARS (Fox et al., 2016). Specifically, we asked participants to complete the sentence: "Whether a given first

⁷The belief that uncertainty can be attenuated by focusing on the predictive value of individual characteristics is a hallmark of relatively more epistemic models of uncertainty.

⁸We selected these attributes first from articles on students' beliefs about the RSL and validated them using interviews with undergraduate students in Harvard's Department of Psychology.

If you had to pick one, which student do you think would end up in a River House?

Student	A	B
In blocking group?	Yes	Yes
Linked?	Yes	No
Blocking group size	8	6
# women	2	5
# varsity athletes	0	2
# legacies	0	2

☒ Student A
☐ Student B

You picked **Student A** to end up in a River House.

Student	A	B
In blocking group?	Yes	Yes
Linked?	Yes	No
Blocking group size	8	6
# women	2	5
# varsity athletes	0	2
# legacies	0	2

What is one thing that **Student B** could change to potentially improve their ability to get into a River House?

Figure 4: Study 2: epistemicness manipulation. Note that the left and right panels appeared on separate pages.

year student will end up in a Quad House or a River House...”. Example items for epistemic uncertainty include “... is knowable in advance, given enough information”, and “is something that well-informed people would agree upon”. Example items for aleatory uncertainty include “... is something that has an element of randomness”, and “could play out in different ways on similar occasions”. This measure yields two subscales (epistemic uncertainty / aleatory uncertainty), each with a theoretical minimum of 1 and a theoretical maximum of 7. Given that the RSL is a random lottery, we expected that focal participants would tend to perceive uncertainty surrounding it as being relatively more aleatory than epistemic.

Choice

To serve as our dependent measure, we count the number of times that each participant selected the predictor who used confidence language (rather than the predictor who used likelihood language). This variable has a theoretical range of [0, 3], with zero indicating that a participant always selected the predictor who used likelihood language and three indicating that the participant always selected the predictor who used confidence language. Recall that focal participants choice of predictor reflects their beliefs about that predictor’s score across all predictions, with predictors’ scores being based on their accuracy, calibration, and resolution.

Given the highly aleatory nature of the RSL, our model would predict that, collapsing across

conditions, focal participants would tend to choose predictors who used likelihood language more often than those who used confidence language. If our manipulations are effective, we would also expect that focal participants in the epistemic condition would perceive uncertainty surrounding the RSL as being relatively more epistemic, compared to participants in the aleatory condition. Therefore, we predicted that participants in the epistemic condition would be less likely to choose predictors who used likelihood language (more likely to choose predictors who used confidence language) compared to participants in the aleatory condition.

Results

Epistemicness

Participants in both conditions perceived the RSL as being highly aleatory ($M_{\text{aleatoryness}} = 5.62$, $SD = 1.06$) and moderately epistemic ($M_{\text{epistemicness}} = 2.93$, $SD = 1.28$). We can then examine whether these beliefs varied as a function of experimental condition, which serves as a manipulation check. First, we find no effect of our manipulation on perceived aleatoryness, $t(521.28) = 0.76$, $p = 0.45$, consistent with the wording of the aleatoryness items alluding to *some* (rather than a large amount of) randomness and/or participants' strong prior beliefs that the RSL is in fact random. We do, however, find an effect on perceived epistemicness. Participants in the epistemic framing condition perceived the RSL as being more epistemic ($M_{\text{epistemicness}} = 3.14$, $SD = 1.02$) than did those in the aleatory framing condition ($M_{\text{epistemicness}} = 2.73$, $SD = 1.09$), $t(521.88) = 3.70$, $p < .001$; Cohen's $d = 0.32$, 95% C.I. for the difference = $[0.15, 0.50]$.

Choice

Consistent with our predictions, participants in both conditions displayed a robust preference for predictors who used likelihood language over those who used confidence language: selecting the predictor who used confidence language only 32.6% of the time. Our central result is that this proportion differed across experimental conditions. Participants in the epistemic framing condition selected the predictor who used confidence language 35.9% of the time (95% C.I. = $[30.9\%, 40.8\%]$) whereas participants in the aleatory framing condition selected the predictor who used confidence language only 29.3% of the time (95% C.I. = $[24.7\%, 33.9\%]$), Cohen's $d = 0.17$, 95% C.I. for the

difference = [0%, 34%].

For our primary analysis, we prespecified a Poisson regression of the specification $count\ confidence = intercept + framing\ condition + grad\ year + error$, with the coefficient of interest being *framing condition*. The result of this test confirms our hypothesis, $\beta = 0.20$, $z(520) = 2.21$, $p = 0.027$. We included *grad year* as a control in an effort to maintain statistical power in the event that responses varied as a function of graduation year.⁹ They did not. Thus, a Chi-squared test yields a qualitatively similar result, $\chi^2(3) = 10.42$, $p = 0.015$.

Robustness

After completing this study and analyzing these results, we learned about some recent concerns that have been raised about inflated false positive rates in Poisson regressions (Ryan, Evers, & Moore, 2021). In short, Poisson regressions assume that the dependent measure has a variance equal to its mean and are particularly sensitive to violations of this assumption. When this assumption is violated, as in our data, Poisson regressions exhibit inflated false positive rates (i.e., more than N% of p values falling below .0N under the null hypothesis.) Fortunately, the false positive rate of any Poisson regression can be precisely estimated using simple simulation-based methods, and the true false positive rate can be used to adjust p-values.

We conducted such a simulation using a publicly available tool developed by Levin (2020). Based on this approach, we estimate our model's true false positive rate to be 10.4% using the $p \leq .05$ threshold. We can use this false positive rate to adjust the p value of our prespecified regression – a process which yields an estimated true p-value 0.066. Details of this process and additional rationale for its use are provided in the supplement.

To further assess the robustness of the result, we fit two additional generalized linear models that address the threat of dispersion in Poisson regressions by modeling it directly. Both yield similar results. The two models are quasi-Poisson: $\beta = 0.20$, $t(520) = 1.84$, $p = 0.067$ and quasi-binomial (Agresti, 2010) $\beta = 0.29$, $t(520) = 1.84$, $p = 0.066$. Note that these analyses produce p-values that are qualitatively identical to the p-value that we estimated via simulation.

⁹For example, if the recency of one's personal experience with the RSL (or the experience of one's classmates) had an effect on beliefs.

Although the strength of evidence from Study 2 is moderate under the analyses described in this section, we interpret the pattern of results as consistent with a true but modestly-sized effect ($d = 0.17$) and inadequate power, for two reasons. First, we note the estimated power of our analyses. For our prespecified analysis, we estimate 63% power to detect an effect of the estimated size. Using the two models that account for dispersion in our dependent measure, we estimate approximately 50% power. All power analyses use the $p = .05$ threshold.

Second, we note that some exploratory analyses suggest that our estimates become more precise when we restrict analyses to participants who were more attentive and/or who had stronger preferences (independent of whether those preferences were consistent or inconsistent with our model.) Notably, we highlight that the context (the RSL), the framing manipulation, and the types of language used by advisors were consistent across all three rounds. As such, our model would predict that participants would display a consistent preference for one language type or another across all three rounds (plus random error.) This is generally the case: 71.6% of participants displayed an unambiguous preference for a single type of language across all three rounds. (Under the null model, this would occur 12.5% of the time.) More specifically and consistent with the ostensibly-random nature of the RSL, 53.3% of participants always selected the predictor who used likelihood language and 18.3% always selected the predictor who used confidence language.

There are several possible explanations for why a participant would make ‘inconsistent’ choices across rounds. The first is random error: participants’ choices were a function of both model-consistent preferences and noise, with all participants’ preferences being translated into choices with similar fidelity. The second is inattentiveness: essentially the same as random error, except participant-variant, such that inattentive participants translate their preferences into choices with less fidelity than do attentive participants. Finally, another explanation is confounding: the possibility that our stimuli vary within-condition on unobserved dimensions that moderate the relationship between focal participants’ models of uncertainty and their choices.

If our stimuli are confounded, all of our analyses are invalid. However, if inconsistent responses are caused by either of the other two explanations, they would be equally likely to err in the hypothesis-consistent direction as in the hypothesis-inconsistent direction. As a result, analyses that are restricted to only those 71.6% of participants who displayed consistent preferences across all three

rounds would produce more accurate estimates of the true treatment effect (while underestimating its standard error.) Restricting our prespecified analysis to only these 376 participants yields a substantially larger *and* more precise estimate, $\beta = 0.45$, $z(371) = 3.69$, $p = 0.0002$, $d = 0.26$. This estimate remains highly significant across all models ($ps < 0.001$) and is robust to imputing the (worst case) standard error from the unrestricted model.

Discussion

Notably, we find a robust and positive main effect of using likelihood language on choice, suggesting that participants in the focal sample perceived predictors who used likelihood language as being more competent than those who used confidence language. This is the pattern that we would expect, given that focal participants perceived the uncertainty surrounding outcomes in the Harvard Housing Lottery to be quite aleatory. This finding is superficially inconsistent with the pattern described by Løhre & Teigen (2024), whose participants perceived others as being more competent when they expressed certainty in “external” (e.g., “it is uncertain”; more similar to our likelihood language) rather than internal (e.g., “I am uncertain”; more similar to our confidence language) terms. We interpret these differences between our findings and theirs as evidence that preferences for others’ language when expressing uncertainty are moderated by one’s own beliefs about uncertainty type.

Despite the Harvard Housing Lottery being explicitly random, our manipulation successfully shifted participants’ perceptions of its epistemicness, and these shifts translated into systematic preferences for different types of uncertainty language. This suggests that the language-competence link operates even when ground truth is knowable – participants who were primed to think of the lottery as more epistemic showed an increased preference for predictors who used confidence language, despite this representing a fundamental misunderstanding of the lottery’s nature.

General discussion

Across two studies, we find evidence that people evaluate others’ competence in part based on the fit between the language those others use to express uncertainty and their own beliefs about the type of uncertainty characterizing the domain. When people perceive outcomes as epistemic –

reducible through additional information and expertise – they prefer advisors who express uncertainty using confidence language (e.g., “I’m 70% sure”). When they perceive outcomes as aleatory – inherently random or unpredictable – they prefer advisors who use likelihood language (“There’s a 70% chance”). This pattern held across diverse domains in Study 1. In Study 2, we causally isolated beliefs about uncertainty type and documented those beliefs’ consequences for making a consequential judgment.

These findings advance our understanding of uncertainty communication in several ways. First, they demonstrate that the confidence-likelihood language distinction documented by Ülkümen et al. (2016) has important consequences for social perception. Our work suggests that people don’t only use different language for different types of uncertainty – they also understand that others do the same and therefore draw nuanced inferences based on others’ choices of language.

Second, our results suggest that competence evaluations incorporate not just accuracy (as has long been demonstrated) or calibration (Levin, 2026) – but also beliefs about the nature of uncertainty itself. Traditional models of advice-taking have focused on advisors’ track records, confidence calibration, and domain expertise. We show that advisors’ implicit models of uncertainty, revealed through language, provide additional diagnostic information. This channel may be particularly valuable when track records are unavailable or when expertise is difficult to verify directly.

Third, we extend the literature on epistemic versus aleatory uncertainty by demonstrating that these abstract concepts have concrete consequences in communication and interpersonal perception. Prior work has shown that uncertainty type affects the weighting of evidence, information search, and credit attribution. We additionally show that it shapes the social dynamics of judgment, influencing whose input we seek and whose advice we trust. This social dimension may amplify individual-level effects: if people preferentially listen to advisors with matching uncertainty models, belief communities may form around shared conceptions of what can be known.

These findings should be interpreted in light of an important limitation. Study 2 did not include a control condition, so we don’t know how people perceive the RSL in the absence of a manipulation, and therefore we don’t know which manipulation is driving the differences we observe (although the manipulation check suggests the epistemicness manipulation is doing more of the work).

Conclusion

When we seek guidance about uncertain outcomes, we face a fundamental challenge. To obtain the best advice, we must identify advisors who not only possess relevant knowledge about the content of their advice, but also who understand the nature of the uncertainty surrounding it.

We show that the language people use to express uncertainty provides a valuable signal of their underlying mental models. Those who say “I’m confident” when confidence is misplaced reveal a basic misunderstanding; those who acknowledge randomness where it exists demonstrate metacognitive sophistication that likely extends to their judgments more broadly.

As decisions increasingly rely on human-algorithm partnerships and distributed expertise, the ability to communicate uncertainty appropriately becomes critical. Our findings suggest that effective uncertainty communication requires more than numerical precision – it demands linguistic choices that signal shared understanding between advisors and decision-makers. By attending to these signals and adapting our own communication accordingly, we can build more effective collaborations in the face of an uncertain world.

References

- Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: Wiley.
- Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57–82. <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>
- Dean of Students Office. (2020). Sophomore lottery. Harvard College.
- DiCara, F. (2020). Why the harvard housing lottery might not be random. Harvard Open Data Project (Medium).
- Fox, C. R., Goedde-Menke, M., & Tannenbaum, D. (2021). *Ambiguity aversion and epistemic uncertainty* (SSRN Scholarly Paper No. 3922716). Social Science Research Network. <https://doi.org/10.2139/ssrn.3922716>
- Fox, C. R., Tannenbaum, D., & Ülkümen, G. (2016). *The empirical case for distinguishing two dimensions of subjective uncertainty*.
- Fox, C. R., Ülkümen, G., Tannenbaum, D., Walters, D. J., & Erner, C. (2022). *Attributions of credit*,

blame, and luck depend on perceived nature of uncertainty.

- Gaertig, C., & Simmons, J. P. (2018). Do people inherently dislike uncertain advice? *Psychological Science*, 29(4), 504–520. <https://doi.org/10.1177/0956797617739369>
- Levin, J. M. (2020). False positive diagnostics: Tools for estimating the true false positive rates of statistical tests. Software tool.
- Levin, J. M. (2026). *When (and why) expressing uncertainty affects perceptions of competence.*
- Levin, J. M., Gilbertson, E., & Broomell, S. B. (2019). *An incentive compatible elicitation of subjective probability density functions.*
- Løhre, E., & Teigen, K. H. (2024). When leaders disclose uncertainty: Effects of expressing internal and external uncertainty about a decision. *Quarterly Journal of Experimental Psychology*, 77(6), 1221–1237. <https://doi.org/10.1177/17470218231204350>
- Ryan, W. H., Evers, E. R. K., & Moore, D. A. (2021). Poisson regressions: A little fishy. *Collabra: Psychology*, 7(1), 27242. <https://doi.org/10.1525/collabra.27242>
- Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2017). Judgment extremity and accuracy under epistemic vs. Aleatory uncertainty. *Management Science*, 63(2), 497–518. <https://doi.org/10.1287/mnsc.2015.2344>
- Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General*, 145(10), 1280–1297. <https://doi.org/10.1037/xge0000202>
- Walters, D. J., Ülkümen, G., Erner, C., Tannenbaum, D., & Fox, C. R. (2020). *Investment behaviors under epistemic and aleatory uncertainty* (SSRN Scholarly Paper No. 3695316). Social Science Research Network. <https://doi.org/10.2139/ssrn.3695316>

Appendix

For reasons described in (Ryan et al., 2021), many Poisson regressions produce p-values that are artificially low, thereby increasing the risk of type I error (false positive results.) One simple way to evaluate whether a given Poisson regression suffers from an inflated false positive rate is to simulate what your statistical tests would look like if you were analyzing your own data under the null hypothesis. The process for doing this with experimental data is very straightforward: simply

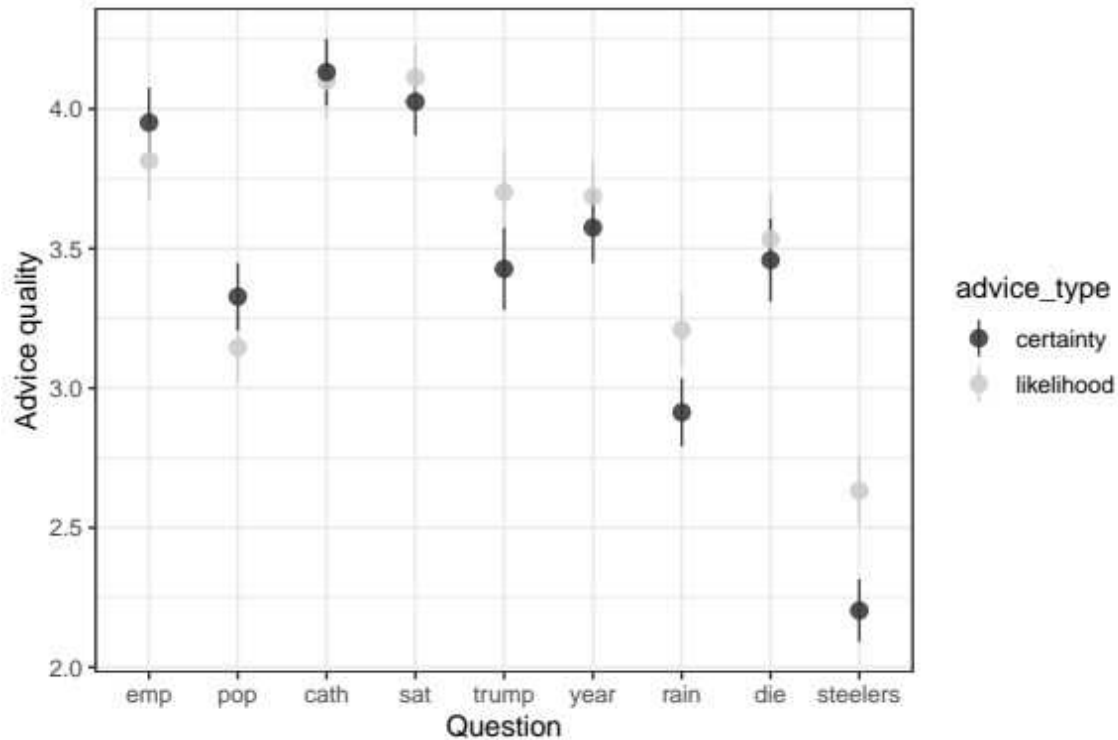


Figure A1: Study 1 Raw Dependent Measure

For each prediction, you'll earn points based on a system that is designed to encourage you to report your true beliefs about first year students' housing outcomes.

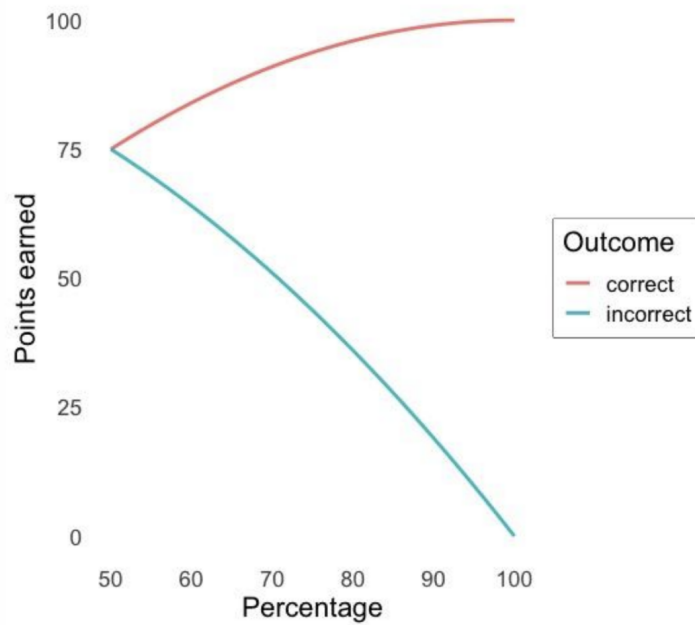
The 10% of participants who earn the most points will each earn a **\$10 cash**.

The graph below illustrates how many points you will earn for each prediction as a function of:

- Whether you are correct or incorrect, and...
- The percentage that you assign to your prediction

If you make an estimate below 50%, it will be subtracted from 100%. For example, points earned for 90% are the same as 10%. 20% is the same as 80%, etc.

Figure A2: Study 2 Predictor Scoring Description



Keep it simple! Just report whichever actual percentage best represents your beliefs.

Figure A3: Study 2 Predictor Scoring Graph

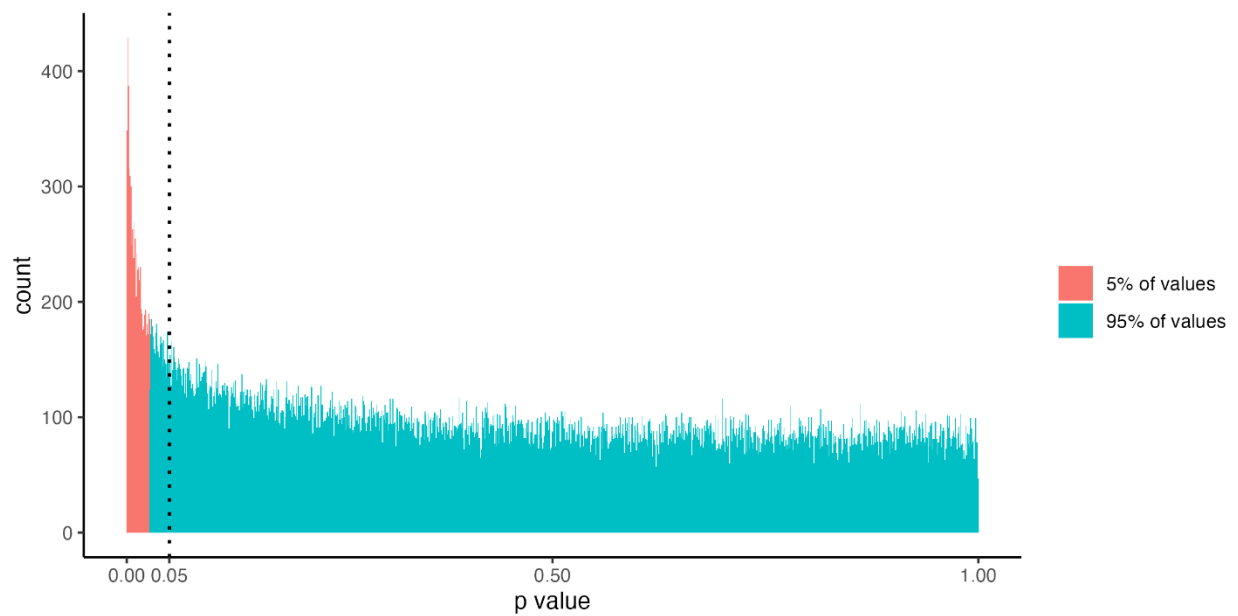


Figure A4: Study 2 Regression Simulations

randomly shuffle the experimental condition variable (such that it should have no association with any measure,) run your regression, and repeat the process thousands of times, each time recording the p-value of the experimental condition variable. If a regression is behaving, you should observe a uniform distribution of p-values, with 5% of values falling at or below .05. If it is not, you will observe something else.

We observe something else. Figure A4 plots the results of 100,000 simulations of this type. Each bin is 0.001 wide (containing 100 regressions on average). From this figure, it is apparent that the distribution of p-values is not uniform, with more p values falling near the bottom of the distribution than the top. Approximately ~10.4% of p-values fall at or below .05. We can reject the hypothesis that this distribution is uniform based on a one-sample Kolmogorov-Smirnov test ($D = 0.090556$, $p < .0001$.)