Abstract

We present novel experimental data pertaining to the use and interpretation of simple probability expressions (such as possible or likely) and complex ones (such as possibly likely or certainly possible) in situations of higher-order uncertainty, i.e., where speakers may be uncertain about the probability of a chance event. The data is used to critically assess a probabilistic pragmatics model in the vein of Rational Speech Act approaches (e.g., Frank and Goodman, 2012; Franke and Jäger, 2016; Goodman and Frank, 2016). The model embeds a simple compositional threshold-semantics for probability expressions, following recent work in formal linguistics (Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2010, to appear; Moss, 2015).

Keywords: pragmatics, communication, rationality, uncertainty, probability

1. Introduction

Although frequently effortless and efficient, communication can be an affair spiked with uncertainty: we talk about things that we do not know for sure many times a day. Will it rain tomorrow? Who will win the next presidential election? What is the risk of an earthquake in this region? Unsurprisingly, natural languages are equipped with devices to communicate uncertain beliefs and the degree of our confidence. We say that it might rain tomorrow; that Trump will probably not win the election again; that an earthquake is certainly unlikely. Statements such as these are mundane and seem unspectacular enough, but on closer look it is a
vexing puzzle what exactly expressions of uncertainty —such as *might, possible, probable, or certainly likely*— mean semantically and how we use and interpret them to communicate uncertain beliefs.

Following Lichtenstein and Newman (1967), much work in experimental psychology has attempted to translate vague uncertainty expressions (mostly *verbal probabilities* like *likely, probable or probably*) into precise numeric chance levels or intervals thereof (see Clark, 1990, for overview). Subsequent studies have attested several sources of contextual effects on the production and interpretation of uncertainty expressions, such as the sensitivity to the prior base rates of the events (Wallsten et al., 1986), the higher variability when expressions are evaluated in context rather than in isolation (Beyth-Marom, 1982; Brun and Teigen, 1988), or the effect of the way in which the space of alternative events is conceptualized (Teigen, 1988; Windschitl and Wells, 1998).

In contrast, philosophers and theoretical linguists have focused on a more abstract description of the semantic meaning contribution of uncertainty expressions, focusing on their logical properties and their contribution to a compositional account of meaning (e.g., Carnap, 1947; Hintikka, 1961; Kripke, 1980; Kratzer, 1977). In her influential approach, Kratzer (1991) gives a uniform semantic treatment of both epistemic modality (e.g., expressions like *might or must*) and verbal probabilities (e.g., expressions like *unlikely or probably*). Kratzer’s proposal is purely qualitative, with no reference to probability measures or similar constructions. However, more recent work convincingly argues for the adoption of a *quantitative* semantic approach which incorporates some reference to probability measures or a similarly rich model-theoretic structure (Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2010, to appear; Moss, 2015). For example, according to the latter approach, a sentence of the form *It is likely that P* is true exactly in those states (or possible worlds) where the probability of event *P* is bigger than a contextually determined threshold $\theta_{\text{likely}}$.

The goal of this paper is to bring closer together experimental and theoretical approaches. We take the above threshold-based semantics for verbal probabilities as our starting point. To capture some of the context-dependent flexibility in the use and interpretation of uncertainty expressions attested in the experimental literature, we turn to a model of pragmatic communication in the tradition of Grice (1975), following recent work on probabilistic pragmatics and Rational Speech Act (RSA) approaches (e.g., Frank and Goodman, 2012; Franke and Jäger, 2016; Goodman and Frank, 2016), in particular the model of Goodman and Stuhlmüller (2013). A major motivation for this choice of modelling approach is the observation that the interpretation of probability expressions seems to be affected by
considering alternative utterances that the speaker could have made but did not, such as in scalar implicature inferences (e.g., Levinson, 1983; Geurts, 2010). For example, an utterance of (1-a) will often suggest that (1-b) is true, because otherwise the speaker would rather have uttered (1-c).

(1) a. The next ball drawn from this urn is probably red.
    b. \( \sim \) It is not certain that the next ball drawn from this urn is red.
    c. The next ball drawn from this urn is certainly red.

The RSA approach offers a convenient framework for modeling the listener’s reasoning about the speaker’s likely choice of alternative utterances and their communicative effects.

Moreover, RSA assumes that the speaker’s choice of messages is governed by the goal to align the speaker’s probabilistic beliefs with those of the listener. This makes RSA particularly useful for modeling the use and interpretation of probability expressions when they are used to communicate even complex uncertain belief states of the speaker.

On top of introducing an RSA model for the use and interpretation of probability expressions, this paper tries to cover new ground in several respects: (i) it considers not only simple but also complex probability expressions (e.g., possibly likely); (ii) it looks at situations of higher-order uncertainty, where speakers may be uncertain about the probability of an event; (iii) it reports on novel experimental data on the production and interpretation of both simple and complex probability expressions, introducing and testing a paradigm that allows the systematic manipulation of higher-order uncertainty; (iv) it critically assesses the predictions of the RSA model in the light of the empirical data, which is used to infer a posteriori credible thresholds that might capture the semantic meaning of probability expressions, given the data and the model of pragmatic use (Schöller and Franke, 2017).

The next section introduces complex probability expressions and higher-order uncertainty. Section 3 then introduces and tests the experimental material used in later parts of this paper to systematically manipulate probabilistic beliefs in experimental participants. Section 4 reports on two experimental studies collecting data on production (4.2) and interpretation (4.3) of simple uncertainty expressions in situations of higher-order uncertainty. We introduce and discuss the details of our pragmatic model (4.4) and scrutinize its predictions in the light of our data. In Section 5 we turn to complex expressions, extend the model accordingly (5.1) and report on two more experimental studies designed to collect data about complex
expressions (5.2). Section 6 critically assesses assets and shortcomings of our modeling and implementation choices and points to relevant future extensions.¹

2. Complex probability expressions & higher-order uncertainty

Complex uncertainty expressions like possibly likely or certainly possible readily occur in written and spoken English, but relatively little attention has been paid to studying their meaning and use. A recent exception is Moss (2015) who considers the following example to demonstrate that we command rather clear intuitions about the appropriateness of the use of complex uncertainty expressions in some cases. Imagine a person called Liem, very fond of green shirts. Liem’s dad Eric has observed Liem wearing green on 500 of 800 consecutive days. Liem’s friend Madeleine made a somewhat similar observation: she observed Liem wearing green 5 times out of 8 consecutive days. The proportion of green observations is exactly the same: 62.5%. However, as Moss argues convincingly, Eric is in the position to assert (2), whereas Madeleine should limit herself to (3):

(2) Liem is definitely likely to be wearing green.
(3) It might be probable that Liem is wearing green.

A conservative compositional analysis of the meaning of (2) or (3) would assume that the outer expression (definitely/might) quantifies the uncertainty of the speaker about whether the inner expression (Liem is likely/probable to wear green) is true. In other words, an intuitively plausible and systematic hypothesis is that speakers of (2) or (3) attempt to pragmatically communicate not only (their point-valued subjective expectation of) the chance of the event Liem wears green, but also their subjective levels of uncertainty surrounding it. We call this uncertainty about uncertainty “higher-order uncertainty.”

A straightforward compositional semantics that relates complex probability expressions with higher-order uncertainty is the influential logic for reasoning

¹The experiments reported in this paper were implemented and run within the psiTurk framework (Gureckis et al., 2016). The code for the experiments, together with the data and the R scripts used for exploration, visualization and analysis, as well as the JAGS implementation of the model are publicly available at https://github.com/mic-he/ProbExp-HOU.
about knowledge and (nested) probabilistic beliefs by Fagin and Halpern (1994).² Fagin and Halpern’s logic is a conservative extension of a modal logic for reasoning about an agent’s knowledge (which we do not need here) and reasoning about an agent’s probabilistic beliefs, including nested probabilistic beliefs about probabilistic beliefs (which is exactly what we need here).³ As the precise technical details are of no relevance here, suffice it to say that Fagin and Halpern’s logic assigns a meaning of the same logical type to any truth-evaluable expression, namely a set of possible worlds, be it a propositional formula or a formula expressing that some agent i has a particular (nested) probabilistic belief. To achieve this, each possible world w in a model for this logic is associated with a valuation function V_w which assigns a truth value (0 or 1) to each atomic proposition letter. Each world w is also associated with a probability measure μ_i,w for every agent i, such that μ_i,w assigns a probability measure to every subset of worlds in the model.⁴ If X is a proposition, i.e., a set of possible worlds, then μ_i,w(X) is the level of credence agent i assigns to X in world w. The resulting semantics for simple and complex probability expressions is straightforward. In the following examples, the a-variant is the sentence to be analyzed, the b-variant a suggestive formal characterization of the denotational meaning (ignoring tense information for simplicity) and the c-variant a gloss of the b-variant in natural language.

(4) a. The next ball drawn from this urn will be red. (= RED)
   b. [RED] = \{w \mid V_w(RED) = 1\}
   c. The set of all worlds in which the next draw is red.

²For application to our experimental scenario later in this paper, the semantics offered by Moss (2015), though finely tuned to explain puzzling intuitions about meaning of and valid reasoning patterns with (nested) probability expressions, is practically equivalent to the more austere logical analysis we endorse here. We chose to stick to the earlier logical analysis because it makes clearer in which way this is a straightforward and conservative account of the meaning of both simple and complex probability expressions. However, nothing of current relevance hinges on this choice.

³The philosophical literature commonly uses a slightly different terminology and would rather speak of an agent’s credence distribution. We occasionally use the terms credence or credence level but stick to what is perhaps the more generally used terminology in other disciplines and so speak of an agent’s probabilistic beliefs or, somewhat sloppily perhaps, an agent’s probability distribution.

⁴Strictly speaking, each world w and agent i are associated with a probability space (Ω_i,w,χ_i,w,μ_i,w) such that Ω_i,w ⊆ W is a subset of possible worlds, χ_i,w is the usual σ-algebra of measurable subsets of Ω_i,w and μ_i,w is a probability measure defined on the elements of χ_i,w. If a set of possible worlds X ⊆ W is not in χ_i,w, we resort to µ_i,w^*(X), where µ^* is the inner measure of μ_i,w as the probability measure that describes agent i’s belief in X at world w.
In sum, this semantics offers an intuitive compositional analysis of complex probability expressions, where (nested) probability expressions are analyzed uniformly as denoting propositions. Higher-order uncertainty affects not only the choice and interpretation of complex probability expressions. It also seems to affect simple uncertainty expressions. Imagine an urn, containing exactly 100 balls of two different colors, red and blue. You know this, but you do not know the exact distribution of colors in the urn. Imagine drawing 8 balls at random from the urn (we will call this “access”) and observing that 5 of them are red (“observation”). Then, after putting the balls back in the urn, (7) seems an appropriate thing to say, whereas (8) does not:

(7) A ball drawn at random from the urn might be red.

(8) A ball drawn at random from the urn will probably be red.

Now imagine the following case for the same urn: you draw 80 balls and count 50 red balls among them. The proportion of observed red balls is the same as before, and yet now (8) is intuitively more appropriate, arguably due to different levels of access. The level of uncertainty about uncertainty seems to affect simple expressions as well. The urn and balls scenario allows us to make the distinction between different levels of uncertainty more precise. It is therefore what this paper will use as an experimental manipulation of speaker’s higher-order uncertain belief states. Knowing the chance of a randomly drawn ball to be red amounts to knowing the distribution of two colors in the urn. This is perfect information, but it corresponds to a first level of uncertainty: until we draw it, we do not know which color the
ball will be. The higher order layer of uncertainty comes into play when we are not sure about the contents of the urn, e.g., when we have only made a partial observation of the urn’s content.

The vast majority of previous research on uncertainty expressions has not explicitly investigated this distinction, let alone precisely manipulated the two levels in an experimental setting. Here we try to do exactly this by leveraging the precision and simplicity of the urn and balls scenario. In light of the controversial question whether, how or when human reasoners deviate from the norms of probabilistic reasoning (e.g., Tversky and Kahnemann, 1974; Gigerenzer and Goldstein, 1996; Jones and Love, 2011; Sanborn and Chater, 2016), the next section investigates whether our urn-based experimental design induces probabilistic beliefs in experimental participants that approximate what we would expect from rational agents.

3. Modeling & manipulating higher-order beliefs

Imagine an urn containing 10 balls of two different colors, red and blue. Any number \( s \in S = \{0, \ldots, 10\} \) of balls in the urn can be red. \( S \) is the state space. The ratio \( s/10 \) is the probability of a randomly drawn ball to be red. Agents typically cannot directly observe \( s \). Instead, they draw a certain number of balls, which we call access and denote it with \( a \). The observed number of red balls is called observation, denoted with \( o \). Obviously, \( o \leq s \) and \( o \leq a \). The ratio \( o/a \) of observed red balls provides a point-valued guess of the probability of drawing a red ball, with higher access values resulting in better approximations. For example, drawing 4 balls from the urn and observing that 3 of them are red (written here as \( 3/4 \)) or drawing 8 and observing that 6 are red (\( 6/8 \)) correspond to the same proportion of observed red balls (75%) but the latter case provides more information: a rational agent will have much more precise beliefs about the contents of the urn after observing 8 balls rather than 4. The goal of this section is to investigate how this intuition can be made mathematically precise enough to enter a formal model of language use and interpretation.

The belief formation of an ideally rational agent who draws \( a \) balls from an urn containing 10 balls and observes that \( o \) are red can be modeled as Bayesian update of the agent’s prior credence or belief distribution over the space of possible quantities of red balls, given the hypergeometric model of the urn (Goodman
and Stuhlmüller, 2013):\(^5\)

\[
\text{rat.bel}(s|o,a) \propto \text{Hypergeometric}(o;a,s,10) \cdot \text{prior}(s)
\]  

(1)

Figure 1 displays the belief distributions computed for the two observations of our running example (conditions \(3/4\) and \(6/8\)), assuming for the time being a flat prior distribution over states. We can see that both distributions have the same mode equal to 8 but the right hand side distribution is more closely concentrated around the mode —it has lower entropy— reflecting the intuition that the agent’s beliefs are more precise.

Equation 1 defines a normative model: it tells us what rational agents should believe about the contents of the urn given a partial observation. The question arises of how good a model this is of actual human behavior, i.e., whether it would be a crude mistake to assume it as the belief formation component of our linguistic model later on. To answer these questions we ran an experimental study to estimate participants’ posterior beliefs about the contents of the urn with a slider bin rating task (e.g., Kao et al., 2014; Degen et al., 2015; Franke et al., 2016).

---

\(^5\)The hypergeometric distribution describes the probability of obtaining a number \(k\) of random draws with a specific feature (successes, e.g. red balls) given a number \(n\) of draws (e.g. a sample of balls from the urn), without replacement, from a finite population of size \(N\) containing \(K\) objects with the wanted feature. The proportionality sign between the two sides of the equation indicates that they are to be equal up to a normalizing constant, which in this case is: \(\sum_{s'=0}^{N} \text{Hypergeometric}(o,a,s',10) \cdot \text{prior}(s')\).
3.1. Experiment 1

Participants. We recruited 104 self-reported native English speakers with IP addresses located in the USA on Amazon’s Mechanical Turk. Participants were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.

Materials and procedure. A brief preliminary phase introduced the instructions of the experiment to the participants and familiarized them with the urn setting. Participants learned that urns contained balls of at most two colors, and they got acquainted with the textual and graphical depictions of possibly partial observations of an urn’s content (see top of Figure 2) and with the input sliders which they would use to report their intuitions (more details below). After this first phase, each participant completed 13 trials. All the trials had the same structure. Each trial was randomly associated with a possibly partial observation of the urn which was not previously selected among the 65 logically possible combinations of access values from 1 to 10 and observation values from 0 to 10. A picture was shown to the participant, representing the selected observation, together with a brief description. For each observation, we asked participants to answer the question “How many red balls do you think there are in the urn in total?” Participants adjusted 11 sliders, one for each possible quantity of red balls in the urn (0, ..., 10). Slider labels ranged from Impossible to Certain, expressing the intuitive likelihood of each quantity having observed the configuration in the stimulus (see Figure 2). We recorded slider ratings as discrete values ranging from 0 (Impossible) to 1 (Certain) with a step of 0.01.

Results. We discarded the answers of 3 participants who had selected Impossible for all the bins in at least one observation condition. For each of the 101 remaining participants we normalized the ratings for each condition, then calculated, for each condition, the average of these normalized ratings across participants (e.g., Kao et al., 2014; Degen et al., 2015; Franke et al., 2016). We thus obtained 65 mean probability distributions, one for each condition, which we take to approximate the central tendency of beliefs held by all participants after the corresponding observation of the urn. As an illustration, Figure 3 displays (in red) the measured distributions corresponding to the fifteen observation conditions which will play a role in the production tasks of Experiment 1 and 2 described in Sections 4 and 5. We notice that the average measured distributions seem to display a reasonable behavior. In particular, the distributions are more and more peaked (towards reasonable values) as the access value increases or, equivalently, the level of higher-order uncertainty decreases.
Our goal is to assess whether the empirically observed slider ratings are compatible with the assumption that participants may have held rational Bayesian beliefs about urn contents. Towards this end, we adopt a proposal for a likelihood function of observed slider ratings by Franke et al. (2016) in a similar setting. Franke et al. (2016) consider a hierarchical probabilistic model of how population-level average beliefs are the central tendency of the individual-level beliefs of all experimental participants, and how each individual’s probabilistic beliefs determine a likelihood of observing a particular slider rating. While Franke et al. (2016) use this model to infer, via Bayesian posterior inference, likely population-level beliefs in domains where no obvious normative belief distribution exists, we are here interested in testing whether the assumption is tenable that the slider ratings we observed could have been produced by individuals who all held the normatively correct Bayesian belief in each condition.

The model is spelled out in detail in Appendix A. The following summarizes its most important ingredients. The model makes the radical assumption that the normative distribution $rat.bel$ defined in Equation 1 above is held by every participant in every condition of our experiment. For a given belief distribution, the model predicts a likelihood of observing a particular slider value. In particular, we assume for simplicity that each slider rating is independent of each other and that
each slider rating is a noise-perturbed realization of the corresponding probability mass prescribed by $\text{rat.bel}$. Concretely, if $r_{i,(o,a),s}$ is the observed rating given by participant $i$ for slider $s$ in condition $(o,a)$, the likelihood is defined by:

$$\logit(r_{i,(o,a),s}) \sim \text{Norm}(\logit(\text{rat.bel}(s|o,a), \kappa), \sigma).$$

(2)

In words, both values $r_{i,(o,a),s}$ and $\text{rat.bel}(s|o,a)$ are mapped by a logit transform from the unit interval to the reals, and we expect the logit-transformed observed slider value to be a realization of the logit-transformed predicted value with normally distributed noise with standard deviation $\sigma$. Additionally, parameter $\kappa$ modulates the steepness of the logit transform of $\text{rat.bel}(s|o,a)$, thereby allowing for the possibility that participants may be affine ($\kappa > 1$) or averse ($\kappa < 1$) to realizing extreme slider ratings close to 0 or 1.

Figure 3: Measured belief distributions (in red) in 15 observation conditions together with posterior predictive distributions (in black). Red ribbons display 95% bootstrapped confidence intervals, grey ribbons display Bayesian 95% highest density intervals of the posterior predictive.

We assessed the viability of the normative belief model in light of experimental data with Bayesian posterior predictive checks (PPCs) (see Section 4.4 for more explanation of PPCs). To do so, we implemented our model in the probabilistic programming language JAGS (Plummer, 2003) and estimated the posterior distribution over parameter values given our data. We collected two chains of 2500
samples from the posterior distributions after an initial burn-in period of 2500 samples. We checked convergence via $\hat{R}$ (Gelman and Rubin, 1992). For each of the 2500 sample vectors of parameter values the model generates a set of posterior predictive distributions for the population level beliefs. Mean values of the posterior distributions are visualized (in black) in Figure 3, together with 95% highest density intervals (HDIs) (Kruschke, 2014). We look at discrepancies between hypothetical and actual data, i.e., points in the plots where the HDIs of the PPCs do not overlap with the confidence intervals of the observed data: in these cases the data is still unexpected or surprising, so to speak, in the light of the model trained on the data. We can observe that in the vast majority of conditions the model can adequately approximate the data, suggesting that the normative model of rational belief formation adopted in the model (and, most importantly, in the pragmatic models presented below) may be a rough but good enough approximation to the population-level belief distributions underlying participants’ choices given partial observations of the urn.

4. Simple uncertainty expressions

4.1. Design

We ran two experimental studies on Amazon’s Mechanical Turk to collect human data pertaining to production (Experiment 2a) and interpretation (Experiment 2b) of simple uncertainty expressions under higher-order uncertainty. The main goal of Experiment 2a was to test whether different levels of higher-order uncertainty, of the kind introduced above, play a role in the production of simple uncertainty expressions. Our intuition is that they do, and that part of the communicative effect of our utterances containing such expressions is indeed the transfer of higher-order uncertain information. If this is so, then it is reasonable to expect that listeners will be able to interpret these utterances accordingly, but it is an empirically open question whether they do: can listeners infer the communicated information about the speaker’s level of higher-order uncertainty? Answering this question was the main goal of Experiment 2b.

Participants in both studies were introduced to the general experimental setting with a short cover story fictitiously describing the experiment as a game in which they would cooperate with another player. The exact description read as follows:

“This experiment is an interactive two player game of chance. The players cooperate to guess the contents of an urn. Both players know that the urn always contains 10 balls of different colors (for example,
red and blue). But only one player (the sender) is allowed to draw a certain number of balls from the urn and look at them. The sender puts the balls back into the urn and gives it a nice shake, then the sender draws a new ball from it. Before looking at it, the sender sends a message to the other player (the receiver). The receiver reads the message and tries to guess the exact contents of the urn.”

Three main elements of our experimental design are summarized in this description. These elements are intended to meet three desiderata for a design in which reasoning about higher-order uncertainty could affect language use (see Section 6 for critical reflection). First, the urn setting and the partial observation procedure allow us to manipulate the level of higher-order uncertainty in a flexible but precise way, which is at the same time intuitive and easy to visualize (see Figure 4 for a sample stimulus). Second, the game-like cooperative scenario prompts participants to reason about the communicative effect of their utterances on other agents (in the production study) and to interpret other agents’ utterances by reasoning about what they could have wanted to communicate (in the interpretation study). Finally, the explicit goal of coordinating to guess the exact contents of the urn allows us to specify that the purpose of the conversation is the transfer of information about the urn (e.g., how many red balls there are in the urn), while at the same time conveying that communicating (or inferring) the level of uncertainty about this information matters.

4.2. Experiment 2a: production

Participants. We recruited 89 self-reported native English speakers with IP addresses located in the USA on Amazon’s Mechanical Turk. Participants were paid
0.75 USD for their participation, amounting to an average hourly wage of approximately 10 USD.

Materials and procedure.. After the initial introductory phase, participants completed a short familiarization phase (which we described to participants as training) in which they took on the role of receivers, reading the messages sent by a (fictitious) sender and trying to estimate the number of red balls contained in the urn. The training consisted of 2 trials in fixed order in which participants were asked to report their guess about the number of red balls contained in the urn, given that the sender had sent, respectively, the messages “The next ball will possibly be red” and “The next ball will certainly be red.”

In the main experimental phase participants played as senders. In each trial, participants were exposed to a picture representing an observation of the urn and were asked to make a prediction about the color of a ball drawn at random from the same urn. The experimental conditions were 15 observations of the urn, as summarized in Table 1. Each fraction in the table represents a possible observation: the denominator is the number of accessed balls, the numerator is the number of red balls observed among them. Our choice of conditions allows us to cover a reasonably wide range of proportions of observed red balls under different levels of higher-order uncertainty. For example, we can realize the same proportion of $\frac{1}{2}$ under three levels of higher-order uncertainty, namely high (corresponding to access equal to 4), low (access=8) and none (access=10).

Each participant completed 9 trials, 3 for each level of higher-order uncertainty, in random order; each trial in each level was randomly associated with one condition which was not previously selected in that level. In each trial participants were told to imagine drawing a certain number of balls from the urn, counting the red balls among them, and putting all the balls back in the urn. In each trial a picture was displayed representing this scenario (see Figure 4). Participants were then asked to make a prediction on the basis of their observation: will a ball drawn at random from the urn be red? Crucially, this prediction must be communicated to the receiver by sending a message of the form

\[
\text{The next ball will [...] be red}
\]

where the gap must be completed by the participant selecting an item from a drop-down menu containing certainly, probably, possibly, probably not and certainly not. The choice of these particular messages was dictated by our desire to cover a reasonably wide range of possibilities, while keeping the task as simple as possible. We measured choice counts in each urn condition.
Table 1: Experimental conditions in Experiment 2a. The fractions are observations of the urn. The labels on the left refer to levels of higher-order uncertainty.

<table>
<thead>
<tr>
<th></th>
<th>0/2</th>
<th>1/4</th>
<th>2/4</th>
<th>3/4</th>
<th>2/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>0/8</td>
<td>2/8</td>
<td>4/8</td>
<td>6/8</td>
<td>8/8</td>
</tr>
<tr>
<td>none</td>
<td>2/10</td>
<td>3/10</td>
<td>5/10</td>
<td>7/10</td>
<td>8/10</td>
</tr>
</tbody>
</table>

Figure 5: Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals (black bars).

Results. Figure 5 displays percentages of participants’ choices of expressions in each observation condition. Strikingly, the proportion of observed red balls over the number of accessed balls seems to have an effect on expression choice. For example, proportion values between 0 and $\frac{1}{3}$ (two leftmost columns in the plot) seem to associate mostly with the expression *probably not*, which is the most frequently chosen expression in all conditions except two (where it is the second most frequently chosen); proportion values of exactly $\frac{1}{2}$ seem to invariably correspond to modal choice of *possibly*; and proportion values between $\frac{3}{4}$ and 1 seem to correspond to modal choice of *probably* (although the pattern is less clear in this case).

We fitted a multinomial logistic regression model with the categorical factor
expression of expression choices as dependent variable and the metric factor proportion of the proportion values corresponding to each condition as predictor. Table 2 summarizes the model. The analysis reveals that proportion has a significant effect on participants’ choices of expression, taking possibly as the reference level. This is an intuitive result, one which we can interpret as a sanity check for our experimental setting.

However, following the intuition that proportion is not all that matters, the main goal of our production task was to collect data in situations of different levels of higher-order uncertainty —here represented by different observations of the urn. Looking at our data from this point of view, we can observe that the same (or close enough) proportion values together with different access values seem to give rise to different expression choices. For example, compare the choices of possibly and possibly in the 3/4 and 6/8 conditions (fourth column from the left, top and middle quadrants in Figure 5): the proportion is the same, a reasonably high 0.75 chance; however, it appears that only the participants who observed this proportion in the lower uncertainty situation (access equal to 8) reliably chose possibly, whereas participants who observed the same proportion but in the higher uncertainty situation (access equal to 4) were almost equally split between probably and possibly. Similar differences can be observed comparing the distributions of expression choices recorded with a proportion of 0 and $a = 2$ or $a = 8$, and similarly with a proportion of 1 and $a = 2$ or $a = 8$.

Multinomial logistic regression reveals that both observation and access values have a significant effect on participants’ choices of expression, taking possibly as reference level. (The model $expression \sim observation + access$ is summarized in Table 3.) Finally, a comparison of the two described models in terms of Akaike’s Information Criterion (AIC) results in a preference for the latter model.

<table>
<thead>
<tr>
<th>expression</th>
<th>estimate</th>
<th>std. error</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>certainly not (intercept)</td>
<td>0.427</td>
<td>0.266</td>
<td>1.607</td>
<td>0.108</td>
</tr>
<tr>
<td>probably not (intercept)</td>
<td>1.462</td>
<td>0.199</td>
<td>7.359</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>probably (intercept)</td>
<td>-2.507</td>
<td>0.303</td>
<td>-8.267</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>certainly (intercept)</td>
<td>-5.430</td>
<td>0.663</td>
<td>-8.185</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>certainly not</td>
<td>-6.439</td>
<td>0.892</td>
<td>-7.218</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>probably not</td>
<td>-4.517</td>
<td>0.478</td>
<td>-9.450</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>probably</td>
<td>3.704</td>
<td>0.443</td>
<td>8.368</td>
<td>&lt;0.001 ***</td>
</tr>
<tr>
<td>certainly</td>
<td>5.665</td>
<td>0.830</td>
<td>6.826</td>
<td>&lt;0.001 ***</td>
</tr>
</tbody>
</table>

Table 2: Multinomial logistic regression on expression with proportion as predictor. Log-Likelihood: $-856.75$, McFadden $R^2 = 0.207$, $\chi^2 = 448.05$, $p < 0.001$. 


We can conclude from this analysis that our manipulation of observation and access played a role in participants’ choices of uncertainty expressions. This result is intuitive, as the access/observation pairs express observations of the urn, which were ultimately the most important (if not the only) sources of information displayed to the participants: it’s plausible to think that they observed the display and formed a belief about the content the urn, on the basis of which they decided which message to send.

However, it is still not clear whether the different levels of higher-order uncertainty induced by the observations directly played a role in the choice or not. It seems possible to argue that even if higher-order uncertainty played a role in the belief formation, maybe participants made their choices without taking the full distributions into account but only a flattened-out, summary value expressing their first-order uncertainty about the proposition \textit{The next ball will be red}. In order to dismiss this interpretation we compared the multinomial logistic model defined above (\textit{expression}~\textit{observation+access}) with two models trying to explain participants’ expression choices on the basis of a single summary value of empirically measured participants’ beliefs (see Section 3). For each condition

\begin{table}[h]
\centering
\begin{tabular}{lccccc}
\hline
 & \textit{estimate} & \textit{std. error} & \textit{t} & \textit{p} \\
\hline
certainly not (intercept) & -2.403 & 0.453 & -5.305 & <0.001 *** \\
probably not (intercept) & -0.919 & 0.258 & -3.568 & <0.001 *** \\
probably (intercept) & -1.959 & 0.289 & -6.783 & <0.001 *** \\
certainly (intercept) & -3.234 & 0.503 & -6.432 & <0.001 *** \\
certainly not (observation) & -1.187 & 0.155 & -7.659 & <0.001 *** \\
probably not (observation) & -0.892 & 0.094 & -9.465 & <0.001 *** \\
probably (observation) & 0.739 & 0.082 & 9.041 & <0.001 *** \\
certainly (observation) & 1.022 & 0.134 & 7.628 & <0.001 *** \\
certainly not (access) & 0.498 & 0.072 & 6.891 & <0.001 *** \\
probably not (access) & 0.427 & 0.047 & 9.039 & <0.001 *** \\
probably (access) & -0.204 & 0.060 & -3.378 & <0.001 *** \\
certainly (access) & -0.436 & 0.113 & -3.872 & <0.001 *** \\
\hline
\end{tabular}
\caption{Multinomial logistic regression on \textit{expression} with \textit{observation} and \textit{access} as predictors. Log-Likelihood: $-783.56$, McFadden $R^2 = 0.275$, $\chi^2 = 594.42$, $p < 0.001$.}
\end{table}

(expression~\textit{observation+access}) despite the added complexity (AIC scores 1729 vs 1591).\footnote{Roughly speaking, AIC estimates the information lost when a particular statistical model is used to represent the data-generating process, taking into account not only the goodness of fit of the model but also its simplicity, e.g. the number of free parameters. This makes AIC especially useful in our setting, as we are comparing models of increasing complexity for the same data.} We can conclude from this analysis that our manipulation of observation and access played a role in participants’ choices of uncertainty expressions. This result is intuitive, as the access/observation pairs express observations of the urn, which were ultimately the most important (if not the only) sources of information displayed to the participants: it’s plausible to think that they observed the display and formed a belief about the content the urn, on the basis of which they decided which message to send.

However, it is still not clear whether the different levels of higher-order uncertainty induced by the observations directly played a role in the choice or not. It seems possible to argue that even if higher-order uncertainty played a role in the belief formation, maybe participants made their choices without taking the full distributions into account but only a flattened-out, summary value expressing their first-order uncertainty about the proposition \textit{The next ball will be red}. In order to dismiss this interpretation we compared the multinomial logistic model defined above (expression~\textit{observation+access}) with two models trying to explain participants’ expression choices on the basis of a single summary value of empirically measured participants’ beliefs (see Section 3). For each condition
we computed the mode and the expected value of the corresponding distribution and fitted two multinomial logistic regression models explaining expression respectively with the metric factors mode and ev. For both of these models, the comparison in terms of AIC resulted again in a preference for the original model with access and observation as summarized in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>obs.+acc. mode ev</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>1591.13 1689.59 1620.96</td>
</tr>
</tbody>
</table>

Table 4: AIC scores of the multinomial logistic regression models expression~observation+access, expression~mode and expression~ev.

Summing up, we have provided evidence that our manipulation of both observation and access had an effect on participants’ choices of simple uncertainty expressions, and that the decision process involved in the choice was likely not limited to summarizing the beliefs induced by the observation of the urn and choosing based on this. From this, we can conclude that different levels of higher-order uncertainty matter for the production of simple uncertainty expressions. But what was the exact role of observation and access in participants’ decision processes? An attempt to answer this question is provided in the form of a computational model of pragmatic language use and interpretation, whose details are spelled out in Section 4.4 below. Before turning to the model, we briefly report on design and results of the interpretation task.

4.3. Experiment 2b: interpretation

Participants. We recruited 145 self-reported native English speakers with IP addresses located in the USA on Amazon’s Mechanical Turk. Participants were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.

Materials and procedure. After the introductory phase, participants completed a short training phase in which they took on the role of senders, making partial observations of the urn and choosing a message to send. In more detail, the training consisted of 3 trials in fixed order, in which participants had to choose an expression among certainly, probably, possibly, probably not and certainly not in response to three partial observations of the urn, respectively 3/6, 1/2 and 3/8.

In the main experimental phase participants played as receiver. The experimental conditions in Experiment 2b coincided with the five expressions that participants could choose from in Experiment 2a (certainly, probably, possibly, probably not and certainly not). We displayed the expressions in the form of messages
sent by the sender. For each expression, participants completed 2 trials of different kinds, in a perfectly balanced design. We alternately recorded participants’ interpretation of the expressions alongside two axes of communicative effect: half of the trials were state trials and recorded participants’ answers to the question “How many red balls do you think there are in the urn? expressed by adjusting a discrete slider ranging from 0 to 10; half of the trials where observation trials and recorded participants’ answers to the questions “How many balls do you think the sender has drawn? And how many of them do you think were red?”, expressed by adjusting two discrete sliders ranging from 0 to 10 (see Figure 6). The choice to split the experiment into state and observation trials was aimed at simplifying each individual trials, adding only some complexity at the level of the whole task.

Results. Figure 7 displays counts of participants’ choices of state, access and observation values in each expression condition. We can observe a number of interesting features of the data. Starting from the choice of state values (i.e., the quantity of red balls in the urn) displayed in the top row of the picture, the guesses of the participants seem to be consistent with expectation: we observe a symmetric behavior of the pairs of basic and negated messages, with certainly not and certainly associated with the extreme values (respectively 0 and 10 red balls in the urn), probably not and probably most frequently associated with 3 and 6-7 red balls, and possibly, exactly in the middle, associated with 5. Moreover, we observe that such symmetry is absent from the counts of access values (middle row). Here we observe another pattern: the distributions associated with certainly and certainly not, instead of being symmetric, appear to be quite similar and the same holds for probably and probably not. In other words, the same expressions are associated with comparable access values, i.e., similar levels of higher-order
Figure 7: Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals (black bars).
uncertainty. The question remains, however, of whether participants playing in the role of receivers are interpreting the messages in accordance with a speaker’s intentions. Can they recover what a speaker is likely to want to communicate with a choice of expression? To answer this question we need to turn to a computational model of language production and interpretation, which would predict how rational senders and receivers should behave in each situation.

4.4. Model

Rational Speech Act. The Rational Speech Act (RSA) approach (e.g., Frank and Goodman, 2012; Franke and Jäger, 2016; Goodman and Frank, 2016) is a probabilistic computational modeling framework in which language production and comprehension are formalized as recursive Bayesian inferences between approximately rational agents. RSA can be seen as a probabilistic formalization of Gricean pragmatics (Grice, 1975; Levinson, 2000), which incorporates insights from decision theory and game-theoretic pragmatics (Benz et al., 2005; Franke, 2017). The probabilistic nature of RSA makes it an especially useful tool for investigating pragmatic phenomena, typically less clear-cut and more fuzzy than, for example, semantic judgments of truth/falsity. This is especially true in situations where agents might not have perfect information, such as those investigated in this paper. The computational nature of RSA models means that they can be implemented and used to make precise quantitative predictions about the modeled phenomena. Therefore, RSA models are explicitly testable against empirical data, making them an ideal tool for investigating the case at hand.

The concrete model proposed here is a conservative extension of the RSA model developed by Goodman and Stuhlmüller (2013), who adopt the partial observation of the urn procedure to model different uncertainty situations in which agents might or might not derive scalar implicatures. Unlike Goodman and Stuhlmüller (2013), we also investigate model predictions and empirical data concerning listeners’ inferences of access, i.e., inferences about the speaker’s higher-order uncertain knowledge state.

Speaker and listener. We model pragmatic communication about the contents of the urn under higher-order uncertainty. The universe of the discourse is the set of natural numbers \( S = \{0, \ldots, 10\} \), where for any \( s \in S \) the proportion \( s/10 \) is the probability that a randomly drawn ball will be red. Mirroring the structure of our experimental setting, we model both speaker behavior and listener behavior. In a nutshell, the speaker is modeled as an approximately rational pragmatic agent who chooses the best message to send to the listener given the situation. The listener
is modeled as a pragmatic reasoner who infers the intended meaning by reasoning about the speaker’s behavior.

In this simplified model of communication, there are three main factors contributing to the speaker’s choice: the speaker’s belief about the world, the literal meaning of the messages, and the goal of communication. First, the speaker’s belief. The speaker draws a certain number of balls from the urn, denoted with $a$ for access, looks at them and counts how many of them are red (this quantity is denoted with $o$ for observation). On the basis of her observation, the speaker forms a rational belief about the contents of the urn, expressed as a discrete probability distribution over $S$. Assuming that the speaker has a prior belief distribution over $S$, the posterior beliefs are defined in Equation 1 above, repeated here as Equation 3:

$$\text{rat.bel}(s|o,a) \propto \text{Hypergeometric}(o|a,s,10) \cdot \text{prior}(s)$$  \hspace{1cm} (3)

From the modeler’s perspective, there is uncertainty about participants’ actual prior over states. A flexible yet manageable representation of modeler uncertainty is required. For convenience, the prior distribution over states is assumed to be a discrete beta-binomial distribution between 0 and 10 with free shape parameters $\alpha_s$ and $\beta_s$ (more about modeler’s priors over $\alpha_s$ and $\beta_s$ below).\(^7\)

Second, the literal meaning. Following the experimental setting, the model assumes that the sender selects a message from the set of alternatives available in the production task. Messages can be formalized as the composition of an uncertainty expression, e.g., probably, with a simple or negated sentence, in this case The next ball will (not) be red, which we abbreviate with ($\neg$)RED. As discussed in Section 1, we follow the logical semantics for (nested) probability expressions of Fagin and Halpern (1994) and combine it with a threshold semantics for uncertainty expression (Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2010, to appear; Moss, 2015). That is, if $X$ is an uncertainty expression and $p$ a simple sentence like RED or $\neg$RED, the meaning of $X(p)$ is the set of worlds $w$ which are associated with a probability measure $\mu_i,w$, where $i$ is the speaker, such that $\mu_i,w(p) > \theta_X$, i.e., the probability of $p$ being true, according to the speaker, is higher than a semantic threshold $\theta_X$ associated with expression $X$. Yet, not every probability of RED is

\(^7\)A beta-binomial distribution describes the distribution of samples from a binomial distribution with parameter $p$, when $p$ is itself sampled from a beta distribution with shape parameters $\alpha$ and $\beta$. The beta-binomial distribution is the conjugate prior of the hypergeometric distribution (Peskun, 2016), which makes it a salient choice in our setting. It is also a convenient parametric distribution over a bounded interval of integers.
equally plausible, given the experimental urn scenario. If agents have knowledge of the urn scenario, only 11 different states of the world are possible: only worlds for which $\mu_{i,w}(\text{RED}) \in \{s/10 \mid s \in S = \{0, \ldots, 10\}\}$ are logically possible. Another way of thinking about this is that our pragmatic model, like other RSA models do too, implicitly encodes a question under discussion (QUD) in the formulation of the state space $S$ against which a given semantic expression is evaluated (Roberts, 2012; Kao et al., 2014; Lassiter and Goodman, online first). In the case here, the QUD for evaluation of a simple probability expression $X(p)$ would be “What is the probability of $p$?” with only eleven logically possible answers in the current context of conversation. These considerations allow us to formulate a simplified semantics of messages in terms of states, i.e., sets of possible worlds $w$ that agree on $\mu_{i,w}$, for the urn-scenario context:

\[
\begin{align*}
[\text{certainly}(p)] &= \{s \in S \mid s/10 > \theta_{\text{certainly}}\} \\
[\text{probably}(p)] &= \{s \in S \mid s/10 > \theta_{\text{probably}}\} \\
[\text{possibly}(p)] &= \{s \in S \mid s/10 > \theta_{\text{possibly}}\}
\end{align*}
\]

For negated sentences we have:

\[
\begin{align*}
[\text{certainly not}(p)] &= \{s \in S \mid s/10 < 1 - \theta_{\text{certainly}}\} \\
[\text{probably not}(p)] &= \{s \in S \mid s/10 < 1 - \theta_{\text{probably}}\}
\end{align*}
\]

Notice that we are not fixing any value for the semantic thresholds a priori. The thresholds are free parameters in the model, whose credible values will be inferred by conditioning on experimental data (Schöller and Franke, 2017).

The literal meaning of the messages allows us to model their communicative effect prior to any pragmatic inference. That is, we assume a purely semantic level of interpretation to ground the recursive process of pragmatic language use and interpretation, as usual in RSA and related models. This level is modeled as an idealized naïve listener who receives a message $m$ and simply updates her prior belief over $S$ on the assumption that $m$ is literally true:

\[
\text{lit.bel}(s|m) \propto \delta_{s \in [m]} \cdot \text{prior}(s)
\]

where the $\delta$ function simply returns 1 if the condition $s \in [m]$ is met (i.e., the message $m$ is true in $s$), and 0 otherwise.\footnote{Notice that we depart with the usual notation of RSA models, which uses $P_{o}(\cdot)$ for the literal listener’s conditional probability of interpretation choice, given a message. We rather follow recent overview articles on RSA (Franke and Jäger, 2016; Goodman and Frank, 2016) and use a less overloaded, more informative notation here, as well as for the speaker’s and pragmatic listener’s choice probabilities later on.} As an illustration, Figure 8 displays
the belief distributions of the naïve listener as a function of the received message, having fixed reasonable values for the threshold parameters and assuming flat priors. This captures basic intuitions about the meaning of uncertainty expressions: the most informative expressions are certainly and symmetrically certainly not, the least informative is possibly, with probably and probably not exhibiting an intermediate (and symmetric) behavior.

Third, the goal of communication. We assume that the speaker chooses her messages aiming to maximize the information transferred to the listener. From the speaker’s perspective, transferring information to the listener amounts to sending messages which bring the listener’s beliefs about the relevant topic (QUD) as close as possible to the speaker’s own beliefs, i.e., which minimize the distance between the two distributions expressing those beliefs. More formally, we compute the expected utility of a message $m$ given an observation of the urn $\langle o, a \rangle$ as the negative Hellinger distance (HD) between the speaker’s belief distribution
given \( \langle o, a \rangle \) and the literal listener’s belief distribution given \( m \):\(^9\)

\[
\text{EU}(m; o, a) = -\text{HD}[	ext{rat.bel}(s|o, a), \text{lit.bel}(s|m)]
\]  

(5)

As an illustration, Table 5 shows the expected utility of messages computed for two partial observations, namely \( \frac{3}{4} \) and \( \frac{6}{8} \): the values in the table are negative HD between each rational belief distribution displayed in Figure 1 and each literal belief distribution displayed in Figure 8. We can observe that the most interesting rows in the table are the ones corresponding to possibly and probably: in both situations these messages have comparatively high EU, but while their EU is very close in the \( \frac{3}{4} \) situation, the EU of possibly drops in favor of probably in the \( \frac{6}{8} \) situation.

<table>
<thead>
<tr>
<th></th>
<th>3 red balls out of 4</th>
<th>6 red balls out of 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>certainly not</td>
<td>-1.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>probably not</td>
<td>-0.91</td>
<td>-1.00</td>
</tr>
<tr>
<td>possibly</td>
<td>-0.46</td>
<td>-0.69</td>
</tr>
<tr>
<td>probably</td>
<td>-0.44</td>
<td>-0.51</td>
</tr>
<tr>
<td>certainly</td>
<td>-1.00</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

Table 5: Examples of EU of each message given two partial observations of the urn, rounded to two decimal places.

The speaker’s behavior depends on the EU of messages: the speaker’s choice probabilities are defined as a softmax function of EU:

\[
\text{speak.prob}(m|o, a) \propto \exp(\lambda \cdot \text{EU}(m; o, a))
\]  

(6)

\(^9\)Goodman and Stuhlmüller (2013) use Kullback-Leibler divergence as a measure of discrepancy between speaker and listener beliefs. This implies that the speaker will never choose a message whose truth she is not absolutely certain of (c.f. Scontras et al., 2018, Chapter 2 and Appendix 2). To see that this is unintuitive, consider an example. Let us naturally assume that for a case with binary outcomes \( \theta \text{probably} > .5 \): for a fair coin we would not say that it will probably land heads next and that it will probably land tails next. Now take the urn scenario where you observe 3 red balls out of 4 balls drawn from an urn holding 10 balls. KL-divergence predicts that the speaker will never say The next ball drawn will probably be red contrary to (our) intuition and contrary to what we see in the data in Figure 5. Hellinger distance is more adequate in our setting because utilities in terms HD allow for pragmatically “true enough” messages to be sent. The Hellinger distance between two discrete distributions \( P \) and \( Q \) is defined as

\[
\text{HD}(P, Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_i (\sqrt{P_i} - \sqrt{Q_i})^2}.
\]
where $\lambda$, sometimes referred to as the “rationality parameter,” is free in the model. As $\lambda$ grows, choice probabilities approach EU-maximization behavior. As an illustration, Figure 9 displays speaker’s probabilities of sending each message given the two partial observations of our running example for $\lambda = 5$. Notice how this non optimized version of the model can already vindicate, at least qualitatively, the intuition (corroborated by our production data as well) that the $6/8$ observation makes the speaker more inclined to say that a randomly drawn ball will probably be red rather than just possibly, whereas the $3/4$ observation does not, even though it corresponds to the same observed proportion (see Figure 5, fourth column).

Our final step is to model a pragmatic listener, who receives a message and interprets it by reasoning about how a pragmatic speaker could have used the message, on the assumption that she has formed a rational belief about the contents of the urn for a particular observation:

\[
\text{listen.prob}(s, o, a | m) \propto \text{speak.prob}(m | o, a) \cdot \text{Hypergeometric}(o | a, s, 10) \cdot \text{prior}(a) \cdot \text{prior}(s).
\]

\[
\text{listen.prob}(s | m) = \sum_{\langle o, a \rangle} \text{listen.prob}(s, o, a | m)
\]

\[
\text{listen.prob}(o, a | m) = \sum_s \text{listen.prob}(s, o, a | m)
\]
where Equations 8 and 9 are obtained from marginalizing the joint distribution defined in 7 respectively over pairs \( \langle o, a \rangle \) and over states \( s \). The listener’s prior over access values prior \( (a) \) from Equation (7) is subject to modeler’s uncertainty. We avail ourselves of the same structure as for the prior over states, prior \( (s) \), namely a beta-binomial prior with free parameters \( \alpha_a \) and \( \beta_a \) (more below). The distributions defined in Equations 6, 8 and 9 allow us to generate the model predictions which we compare to the experimental data collected respectively in the production task and the interpretation task of Experiment 1.

**Model evaluation and criticism.** First, we used the experimental data to infer credible values for the free parameters of the model, i.e., the shape parameters \( \alpha_s, \beta_s \) and \( \alpha_a, \beta_a \) of the beta-binomial models of participants’ prior over states and access values, respectively; the semantic thresholds \( \theta_{\text{certainly}}, \theta_{\text{probably}} \) and \( \theta_{\text{possibly}} \) and the rationality parameter \( \lambda \). To do so we implemented the model in the probabilistic programming language JAGS (Plummer, 2003) and estimated the posterior distribution over parameter values given our data. In more detail, the model assumes that the observed counts of expression choices, of state, access and observation values in each condition are samples from multinomial distributions with weights equal to the probabilities predicted by the model (i.e., the functions \( \text{speak.prob} \) and \( \text{listen.prob} \)) in the corresponding condition.

We remained relatively uncommitted with respect to the prior distributions over parameter values, assuming flat distributions with support \([0; 1]\) for the thresholds and \([0; 20]\) for \( \lambda \); the prior distribution over states and access values was defined for convenience as a beta-binomial distribution, parametrized in terms of mode \( \omega \) concentration \( \kappa \) (Kruschke, 2014):

\[
\begin{align*}
\theta_{\text{certainly}} & \sim U(0, 1) \\
\theta_{\text{probably}} & \sim U(0, 1) \\
\theta_{\text{possibly}} & \sim U(0, 1) \\
\lambda & \sim U(0, 20) \\
\kappa_{s,a} & \sim \text{Gamma}(0.01, 0.01) \\
\omega_{s,a} & \sim U(0, 1) \\
\alpha_{s,a} & = \omega_{s,a} \cdot (\kappa_{s,a} - 2) + 1 \\
\beta_{s,a} & = (1 - \omega_{s,a}) \cdot (\kappa_{s,a} - 2) + 1 \\
prior(s) & = \text{Betabinom}(s|\alpha_s, \beta_s, 10) \\
prior(a) & = \text{Betabinom}(a|\alpha_a, \beta_a, 10)
\end{align*}
\]

Figure 10 displays a representation of the full data-generating model as a probabilistic graphical model (Lee and Wagenmakers, 2014). We collected two chains of 2500 samples from the posterior distributions after the initial burn-in period of 2500 samples. We checked convergence via \( \hat{R} \) (Gelman and Rubin, 1992). Each sample is a vector containing one inferred value for each parameter.

Table 6 summarizes the results for the semantic threshold parameters \( \theta_{\text{certainly}}, \theta_{\text{probably}}, \theta_{\text{possibly}} \) in terms of mean inferred values and 95% highest density inter-
Figure 10: The data-generating model. White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence.
vals (HDIs). Figure 11 display posterior cumulative density distributions for the threshold parameters. These results showcase a nice feature of the model, namely that it does not assume the values for the semantic thresholds from the beginning but it is nonetheless able to infer plausible and intuitive values given the data.\footnote{The results for the remaining free parameters are summarized in the following table (where we report only $\alpha$ and $\beta$ for the beta-binomial distributions):}

\begin{table}[h]
\centering
\begin{tabular}{l|cccc}
\hline
 & $\theta_{\text{possibly}}$ & $\theta_{\text{probably}}$ & $\theta_{\text{certainly}}$ \\
\hline
lower & 0.200 & 0.500 & 0.904 \\
mean & 0.247 & 0.549 & 0.949 \\
upper & 0.299 & 0.594 & 1.000 \\
\hline
\end{tabular}
\caption{Mean inferred values and HDIs (in terms of lower and upper boundary) of the semantic threshold parameters free in the model, given experimental data collected in Experiment 1.}
\end{table}

To assess model quality, we look at samples of hypothetical repeat-data $D_{\text{rep}}$ from the posterior predictive distribution (where $\theta$ is the vector of all free model
parameters):

\[ P(D_{\text{rep}} \mid D_{\text{obs}}) = \int P(\theta \mid D_{\text{obs}}) \cdot P(D_{\text{rep}} \mid \theta) \, d\theta \]

A posterior predictive sample \( D_{\text{rep}} \) is obtained by taking a sample \( \theta \sim P(\theta \mid D_{\text{obs}}) \) from the posterior distribution over model parameters given the observed data \( D_{\text{obs}} \), and then sampling a likely hypothetical alternative data point \( D_{\text{rep}} \sim P(D_{\text{rep}} \mid \theta) \) from the model’s likelihood function for parameters \( \theta \) (Gelman et al., 2014; Kruschke, 2014). For each of our 2500 samples from the posterior over \( \theta \), we generated one \( D_{\text{rep}} \) for each of the three relevant observations: speaker choices of expression (Equation 6), listener choice of state (Equation 8) and listener choice of access-observation pairs (Equation 9). In order to get an overall evaluation of the model we correlated each set of predictions with the corresponding set of experimental data, collecting the results in vectors of Pearson’s correlation score. Table 7 summarizes the results in terms of mean correlation scores and HDIs. We observe that all the four means and HDIs are reassuringly high, suggesting that the model was overall able to capture regularities in the data.

<table>
<thead>
<tr>
<th></th>
<th>expression</th>
<th>state</th>
<th>access</th>
<th>observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower</td>
<td>0.824</td>
<td>0.666</td>
<td>0.736</td>
<td>0.766</td>
</tr>
<tr>
<td>mean</td>
<td>0.861</td>
<td>0.741</td>
<td>0.798</td>
<td>0.819</td>
</tr>
<tr>
<td>upper</td>
<td>0.902</td>
<td>0.824</td>
<td>0.858</td>
<td>0.868</td>
</tr>
</tbody>
</table>

Table 7: Mean Pearson’s correlation scores and HDIs between model posterior predictive distributions and experimental data collected in Experiment 1.

Bayesian data analysis allows us to supplement correlation scores with a more detailed comparison between model predictions and experimental data via posterior predictive checks (PPCs), displayed in Figures 12 and 13. In particular we look at discrepancies between hypothetical and actual data, i.e., points in the plots where the “confidence” areas do not overlap with the diagonal: in these cases the observed data is still unexpected or surprising, so to speak, in the light of the model trained on the data. In other words, the model fails to predict those data points. It is to be expected that the model will “fail” a PPC for some conditions due to multiple comparisons alone. What matters most is whether there are theoretically insightful patterns of systematic failure that might point to substantial conceptual shortcoming of the model.

Looking at the production data (Figure 12), we can observe that the patterns displayed by the data seem to be captured relatively well by the model. The most
Figure 12: Percentages of expression choices in each observation condition compared to posterior predictive distributions. The rectangular “confidence” areas are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive.

Figure 13: Counts of state, access and observation values choices in each expression condition compared to posterior predictive distributions. The rectangular “confidence” areas are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive.
frequently chosen expression in each observation condition is always correctly predicted by the model (with the exception of one case: in the $0/2$ condition the model underpredicts *possibly* and overpredicts *probably not*). In the majority of conditions the model correctly predicts the second most frequently chosen expression too. Looking at the most glaring discrepancies in the plot we can observe that the model tends to underpredict *possibly* when the proportion is equal to $1/2$ (middle column, especially third and second row). Moreover, the model underpredicts *probably not* in favor of *possibly* with low proportions and no higher-order uncertainty (i.e., $o = 2$ or $o = 3$ and $a = 10$, bottom left corner); and symmetrically *probably* is underpredicted in favor of *possibly* with high proportions and no higher-order uncertainty (i.e., $o = 7$ or $o = 8$ and $a = 10$, bottom right corner). In general, these observations seem to point to a model which is a little more conservative or cautious, so to speak, than the participants. Turning to the interpretation data displayed in Figure e13, PPCs show that the patterns displayed in the data are captured quite well by the model, with the exception of a number of discrepancies, generally where the model seems once again to be more cautious than the participants. All in all, a detailed comparison of the predictions of the model to the observed data in each experimental condition does not reveal any obvious systematic failure of the model. It is not the case, for instance, that any particular message is consistently predicted to occur with a higher probability than attested in the data.\footnote{For example, although as pointed out by a reviewer the phrase *The next ball will possibly be red* may sound less natural than comparable sentences with other probability expressions, a model which does not assume any differences in the speaker’s baseline preference for messages does not overpredict choice rates for *possibly*.

Interim summary.} The experimental data suggests that subtle manipulations of higher-order uncertainty affect production choices of simple uncertainty expressions. Likewise, listeners appear to draw systematic inferences about a speaker’s higher-order uncertain belief state, assigning different levels of credence to access values after observing different simple uncertainty expressions. A model of goal-oriented pragmatic communication was formulated, revolving around the assumptions that interlocutors hold rational higher-order beliefs about the urn scenario and aspire to communicate these complex beliefs. Conditioning on the empirical data, the model recovered values of latent semantic threshold parameters which are intuitive and in line with the relevant literature. Model criticism revealed, as can generally be expected, some mismatches between posterior model predic-
tions and observed data, but no obvious systematic failure to capture particular patterns. We conclude that participants can reason about higher-order uncertainty in this communication scenario, even with simple uncertainty expressions, and they do so, in approximation, in line with a model of rational belief formation and goal-oriented communication. The question we turn to next is whether similar conclusions hold for complex uncertainty expressions as well.

5. Complex uncertainty expressions

The main challenge faced when extending the model of the previous section is how to include complex uncertainty expressions. Not all technical solutions are conceptually equally plausible. Here, we would like to explore what is perhaps the most conservative way of reconciling a lean compositional semantic analysis of complex expressions with a key assumption about pragmatic language use inherent in RSA models, namely that speakers choose expressions based on how well they will help align a literal interpreter’s beliefs with their own.

5.1. Model

The basic setup of the model is the same as before, with the exception of the set of expressions available to the speaker, which now contains 3 simple expressions, i.e., likely, possible, unlikely together with 9 complex expressions obtained by combining the simple ones with 3 modifiers, i.e., certainly, probably, might be. The speaker sends messages of the form It (is) [...] that the next ball will be red, where the gap is to be filled with the expressions in Table 8. The choice of these particular complex messages was dictated by our desire to cover a reasonably wide range of possibilities in a balanced way: the three simple expressions are on a scale from unlikely to likely and each of them appears nested under each of the modifiers, which can in turn be placed on a scale from might to certainly. According to our intuitions, the resulting 9 complex messages vary with respect to their naturalness, but they are all grammatical. An informal search on the Hansard corpus, containing transcriptions of speeches given in the British parliament from 1803-2005, essentially confirmed our intuitions. For example, we go from ~20 occurrences of might be unlikely and probably possible to ~200 of certainly possible and might be likely, up to ~6700 of might be possible.\(^\text{12}\)

There are three main interrelated differences between the simple model and the complex one. The first, and perhaps most obvious, is that we need a specification

\(^{12}\text{Data as of April 25, 2018 retrieved from https://www.hansard-corpus.org.}\)
likely possible unlikely
certainly likely certainly possible certainly unlikely
probably likely probably possible probably unlikely
might be likely might be possible might be unlikely

Table 8: Complex expressions.

of the literal meaning of complex messages. Towards a compositional analysis, we first consider the semantics of simple messages, which are just as in the previous section and model:

\[
\begin{align*}
\text{[likely}(p)\text{]} &= \{s \in S \mid s/10 > \theta_{\text{likely}}\} \\
\text{[possible}(p)\text{]} &= \{s \in S \mid s/10 > \theta_{\text{possible}}\} \\
\text{[unlikely}(p)\text{]} &= \{s \in S \mid s/10 < 1 - \theta_{\text{likely}}\}
\end{align*}
\]

where \(\theta_{\text{likely}}\) and \(\theta_{\text{possible}}\) are free parameters in the model, as before.\(^{13}\)

By the logical semantics of Fagin and Halpern (1994), the denotation of a complex expression \(Y(X(\text{RED}))\), where \(X\) and \(Y\) are uncertainty expressions and \(\text{RED}\) is the proposition \(\text{The next draw will be red}\), is the set of worlds \(w\) where the speaker assigns a probability higher than \(\theta_{Y}\) to the proposition \(X(\text{RED})\) (see example (6)). Similar to our previous considerations relating to simple expressions, the urn-based scenario also imposes natural constraints on the set of possible interpretations of complex uncertainty expressions: not every probabilistic belief about \(X(\text{RED})\) is compatible with a rational belief obtained from a partial or complete observation of the urn’s contents. Consequently, as before, we can use a simpler contextually-restricted semantics for complex expressions in terms of those distinctions between types of possible worlds that the urn scenario makes relevant, which can here be identified with the sets of pairs \(\langle o, a \rangle\) of observations \(o\) and accessed balls \(a\), as these straightforwardly yield the corresponding relevant beliefs

\[^{13}\text{Notice that we assume here that unlikely is essentially interpreted as the logical negation of likely. This assumption, though not uncontroversial, is compatible with the empirical results of Tessler and Franke (2018) who found that expressions like unhappy are interpreted like not happy when presented in isolation, i.e., when a speaker only utters either one of them. Only when listeners interpret multiple utterances from the same speaker, also including expressions like not unhappy, the interpretation assigned to unhappy is more negative than that of not happy. Tessler and Franke predict these empirical results with an RSA model that includes the listener’s uncertain reasoning about the speaker’s likely interpretation of negation markers like un.- The present model does not include this potential level of listener uncertainty. This is presently a mere practical choice in order to keep the complexity of the model manageable.}\]
about $X(\text{RED})$ that a rational speaker might hold in this context:

$$[[Y(X(\text{RED}))]] = \{ \langle o, a \rangle \mid \sum_{s \in [X(\text{RED})]} \text{rat.bel}(s|o, a) > \theta_Y \} \quad (10)$$

To summarize in intuitive terms, the complex expression $Y(X(\text{RED}))$ denotes the set of all higher-order probabilistic beliefs—which are here conveniently identifiable as the set of all observation-access pairs—about the probability of RED where the probability of $X(\text{RED})$—which is a statement about the probability of RED—exceeds $\theta_Y$.

Based on these semantics, we can define how literal listeners will update their contextually relevant prior beliefs after hearing a simple or complex probability expression. Generally, literal listeners update their prior beliefs by ruling out possible worlds that are incompatible with the semantic meaning of the observed expression. For the purposes of modeling later experimental data, we once more formulate the literal listener’s posterior beliefs in terms of state distinctions that are relevant to the given urn scenario. Concretely, the denotation of a simple expression $X(\text{RED})$ is a set of probabilities assigned to the truth of RED, i.e., a set of states $[[X(\text{RED})]] \subseteq S$, whereas the denotation of a complex expression $Y(X(\text{RED}))$ is a set of observation-access pairs, i.e., $[[Y(X(\text{RED}))]] \subseteq O \times A$. If prior($s$) and prior($a$) are the literal listener’s priors over states and access values, the priors over observation-access pairs are:

$$\text{priors}(o, a) = \text{priors}(a) \cdot \sum_s \text{priors}(s) \cdot \text{Hypergeometric}(o \mid a, s, 10) \quad (11)$$

Consequently, the literal listener’s posterior beliefs after observing a message can be conveniently represented for the urn context as:

$$\text{lit.bel}(s \mid X(\text{RED})) \propto \delta_{s \in [[X(\text{RED})]]} \cdot \text{prior}(s)$$
$$\text{lit.bel}(o, a \mid Y(X(\text{RED}))) \propto \delta_{(o, a) \in [[Y(X(\text{RED}))]]} \cdot \text{prior}(o, a) \quad (12)$$

Notice that, on one level, the interpretations of simple and complex expressions are entirely parallel: both rule out semantically incompatible possible worlds from the literal listener’s beliefs. Yet, on another level, there is a difference in their interpretation. Since simple expressions semantically denote “simple uncertainty,” so to speak, while complex expressions denote higher-order uncertainty, the literal listener’s beliefs reflect this distinction in terms of the state distinctions which are relevant for the interpretation of this or that kind of semantic information.
With these literal interpretations, how should we uniformly model the speaker’s conversational goal that governs her choice of utterance? Here we assume, conservatively, that the speaker will still try to minimize the distance between her belief about the conversationally relevant probability of RED to the belief the literal listener holds about the probability of RED after hearing a message. In the case of simple expressions, the case is straightforward (like in Section 4.4). The speaker holds a second-order probabilistic belief about the probability of RED and the literal listener does too. So we define the expected utility of a simple message in terms of the distance between these two second-order beliefs. In contrast, a complex expression induces in the literal listener, by virtue of its compositional semantics, a third-order probabilistic belief, namely a probability distribution over second-order speaker belief states. In that case, a natural and conservative extension is to define the expected utility of a complex message in terms of the expectation of the distance between relevant second-order beliefs under the literal listener’s third-order beliefs. In other words, we may imagine the literal listener to sample an interpretation, i.e., a second-order speaker belief, with a probability given by the third-order belief induced by a complex probability expression. The utility resulting from such an interpretation choice is then just the distance between the two relevant second-order beliefs. This results in the following definition:

$$\text{EU}(m; o, a) = -\sum_{\omega} \text{lit.bel.alt}(\omega | m) \cdot \text{HD}(\text{rat.bel}(\cdot | o, a), \omega)$$

where $\omega$ ranges over the set of relevant second-order beliefs about the probability of RED and $\text{lit.bel.alt}(\omega | m)$ is just a convenient alternative notation for the literal listener’s beliefs which uniformly represents beliefs induced by simple and complex expressions as (possibly degenerate) third-order distributions:

$$\text{lit.bel.alt}(\omega | m) = \begin{cases} 1 & \text{if } m \text{ is simple and } \omega = \text{lit.bel}(\cdot | m) \\ \text{lit.bel}(o, a | m) & \text{if } m \text{ is complex and there is an } a \text{ and } o \\ & \text{such that } \omega = \text{rat.bel}(\cdot | o, a) \end{cases}$$

Having defined the EU of each message given each observation we can finally turn to modeling the pragmatic speaker’s and listener’s behavior. Here the complex model does not diverge in any way from the simple one:

$$\text{speak.prob}(m|o, a) \propto \exp(\lambda \cdot \text{EU}(m; o, a))$$
\[
\text{listen.prob}(s, o, a|m) = \text{speak.prob}(m|o, a) \cdot \text{Hypergeometric}(o|a, s, 10) \cdot \text{prior}(a) \cdot \text{prior}(s)
\]  
(16)

As before, we derive \(\text{listen.prob}(s|m)\) and \(\text{listen.prob}(o, a|m)\) from Equation 16 by marginalization.

5.2. Experiment 3

Participants. We recruited 255 self-reported native English speakers with IP addresses located in the USA on Amazon’s Mechanical Turk. 104 participants took part in the production task and 151 participants took part in the interpretation task. Participants were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.

Materials and procedure. For the most part, the experiment had the same structure and content as Experiment 2. Participants read the same cover story of Experiment 2 and they completed similar training phases.\(^{14}\) The experimental conditions of the production task were the same 15 observations summarized in Table 1. Participants in the production task completed 12 trials, one for each of 12 unique conditions in random order. In each trial participants looked at the sequence of pictures corresponding to the selected condition (see Figure 4) and made a prediction about the color of a randomly drawn ball. As per Experiment 2, the prediction must be expressed by completing a message which they would send to the receiver. In this case, the message had the form

\[
\text{It [...] [...] the next ball will be red}
\]

where the gaps had to be filled with the most appropriate combination of auxiliary/modifier and simple uncertainty expressions selected from two drop-down menus (Figure 14). The experimental conditions for the interpretation task were the 12 message combinations obtained combining the 4 outer expressions \(\text{is, is}
\]

\(^{14}\)In more detail, participants of the production task completed a training in which they played 2 fixed rounds in the role of receivers, reporting their intuitions about the content of red balls in the urn having received, respectively, the message “It’s possible that the next ball will be red” and “It’s certainly likely that the next ball will be red.” Participants of the interpretation task played 3 fixed rounds as sender, observing the conditions \(3/6, 1/2\) and \(3/8\), respectively, and choosing a message to send (see main text for the options).
probably, might be, is certainly with the 3 inner expressions possible, likely, unlikely.\footnotemark[15] Participants completed 24 trials, alternating state trials and observation trials for each of the 12 expressions in random order. The recorded measures were the same as in Experiment 2.

Results. We discarded the answers given by 2 participants in the production study who explicitly admitted a poor understanding of the task. Moreover, we discarded the answers given by one participant in the interpretation task who selected both access and observation values equal to 0 for at least one expression.\footnotemark[16]

Figure 15 shows percentages of 102 participants’ expression choices in each observation condition. The data is interesting in a number of respects. We would like to highlight three. A first basic but important observation is that participants selected also complex expressions in a systematic manner. Five out of the nine complex expressions are the modal choice in at least one experimental condition, as are all three simple expressions. Secondly, by visual inspection, the observed data seems to follow a pattern that is in line with the general predictions of the pragmatic model. The middle column of Figure 15 represents situations in which the speaker sees an equal number of red and blue balls. In these situations $s = 5$ is the most likely state under the relevant rational beliefs (with unbiased priors). We would therefore expect to see frequent choices of expressions which include

\footnotetext[15]{The choices in the second drop-down menu also included probable. This was to offer participants their favorite pick between probable and likely. To keep matters simple, all analyses and modeling in this paper treat probable and likely in inner position as synonymous and choices of either as belonging to the same category, which we will simply refer to as a choice of likely. This is in line with usual assumptions in the literature, although it is worth noticing that recent work found experimental evidence of subtle differences between probable and likely (Lassiter and Goodman, 2015).}

\footnotetext[16]{While such a selection was clearly logically possible, it explicitly contradicted the instruction given to the participants.}
Figure 15: Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals (black bars).
Figure 16: Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals (black bars).
possible. In the two columns on the right we find epistemic states where a higher proportion of red balls is subjectively more likely than a lower. We therefore expect more choices of expressions with likely. The reverse is the case in the two columns on the left, where we expect more choices with unlikely. These general regularities indeed show in the data. At the same time, going through Figure 15 row-wise from top to bottom, speakers have increasing access, so less higher-order uncertainty. With less higher-order uncertainty we see a general trend of increasing use of certainly or is in connection with unlikely (in the two left-most columns) and likely (in the two right-most columns). Finally, an interesting observation relates to the choice between of certainly and is, e.g., in conditions with access 8 or 10. In a sense, certainly appears to be an intensifier, compared to is. Choice counts of certainly likely increase from condition 6/8 to 8/8 and from 7/10 to 8/10, whereas choice counts of is likely decrease. Similarly, choice counts of certainly unlikely increase from condition 2/8 to 0/8 and from 3/10 to 2/10, whereas choice counts of is unlikely decrease. In a similar sense, might be might be something like a “downtoner.” These patterns are also reflected in the interpretation data, to which we turn next, and it will be interesting to see whether the model captures them.

Figure 16 shows counts of 150 participants’ choices of state, access and observation values in each expression condition. Our model predicts that interpretation, by Bayes rule, follows the likelihood of production choices, so that we should expect (ignoring strong prior effects) to see patterns similar to those observed in the production data. This is indeed what we find. Modal and mean interpretation choices of state s are lowest for expression choices with unlikely, higher for possible and highest for expressions with likely. Moreover, participants’ interpretation choices for the number of observed red balls are very well behaved and appear to line up with the interpretation of the state component under a rational belief model.

As in the production data, we see symptoms of something like “intensifying/downtoning effects” of outer expressions certainly and might be in the mean interpretation choices for the state and access dimension, plotted in Figure 17. For example, listeners estimate the true number of red balls to be higher when they hear certainly likely (mean interpretation of state 7.05 and bootstrapped 95% CI [6.76; 7.32]) than when they hear is likely (mean 6.39, [6.15; 6.65]). If we compare the first to the last row (especially is (un-)likely to might be (un-)likely) we see a tendency to a similar “downtoning effect.” For example, the interpretation of might be likely gets a mean of 5.35 ([5.13; 5.59]).

Finally, the interpretation along the access-dimension is also interesting be-
cause it shows how well informed or knowledgeable the speaker is estimated to be after hearing certain messages. Here, too, we see a trend to estimate the speaker to be more informed (higher access) when an expression is modified with *certainly* than when copula *is* is used, and even less competent when *might be* is used. For example, mean interpretation of access for *certainly likely* is 6.52 ([6.3; 6.76]), for *is likely* it is 6.15 ([5.91; 6.4]) and for *might be likely* it is 5.54 ([5.3; 5.8]).

**Model evaluation and criticism.** We adopted the same procedure as described in Section 4.4. First, we inferred credible values for the free parameters of the model given the data. To ensure convergence of the more complex model, we ran 2 chains with 5000 steps each and a burn-in of 4000, resulting in 2000 samples in total. The results obtained for the semantic threshold parameters are summarized in Table 9. Notice that the values for semantic thresholds inferred here are not credibly different from those inferred for the corresponding expressions based on the simpler model and data from Section 4.4, reported in Table 6: the 95% HDIs of the earlier inference overlap those of the current inference for all three

---

<table>
<thead>
<tr>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>4.639</td>
</tr>
<tr>
<td>$\alpha_s$</td>
<td>4.932</td>
</tr>
<tr>
<td>$\beta_s$</td>
<td>4.882</td>
</tr>
<tr>
<td>$\alpha_a$</td>
<td>27.956</td>
</tr>
<tr>
<td>$\beta_a$</td>
<td>20.379</td>
</tr>
</tbody>
</table>

---

17 The results for the remaining free parameters are summarized in the following table:
threshold parameters. This is reassuring. Our more complex model, trained on a different set of empirical data, can recover very similar values for the latent semantic thresholds to the ones recovered by the simple model. On top of that, it is interesting to see that modal expressions in outer position are estimated to have a higher semantic threshold than related expressions in inner position. E.g., the threshold of might be (a nesting expression) is estimated to be higher than that of possible (a nested expression) and similarly for probably and likely.

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>mean</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>θpossible</td>
<td>0.200</td>
<td>0.251</td>
<td>0.295</td>
</tr>
<tr>
<td>θmight</td>
<td>0.328</td>
<td>0.332</td>
<td>0.336</td>
</tr>
<tr>
<td>θlikely</td>
<td>0.504</td>
<td>0.549</td>
<td>0.599</td>
</tr>
<tr>
<td>θprobably</td>
<td>0.629</td>
<td>0.690</td>
<td>0.745</td>
</tr>
<tr>
<td>θcertainly</td>
<td>0.969</td>
<td>0.982</td>
<td>0.988</td>
</tr>
</tbody>
</table>

Table 9: Mean values and HDIs of inferred values for the semantic threshold parameters given experimental data from Experiment 2.

Second, we computed a posteriori mean credible correlation scores between model predictions and experimental data, as summarized in Table 10. The correlations between interpretation data and the posterior predictions of the listener model are reassuringly high. On the other hand, the correlation between the data obtained in the production task and the posterior predictions of the speaker model is noticeably lower, also in comparison to the results from the simple expression model of Section 4.

<table>
<thead>
<tr>
<th></th>
<th>expression</th>
<th>state</th>
<th>access</th>
<th>observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower</td>
<td>0.596</td>
<td>0.812</td>
<td>0.818</td>
<td>0.915</td>
</tr>
<tr>
<td>mean</td>
<td>0.654</td>
<td>0.849</td>
<td>0.853</td>
<td>0.934</td>
</tr>
<tr>
<td>upper</td>
<td>0.719</td>
<td>0.883</td>
<td>0.888</td>
<td>0.952</td>
</tr>
</tbody>
</table>

Table 10: Mean Pearson’s correlation scores and HDIs between model posterior predictive distributions and experimental data collected in Experiment 2.

Overall correlation scores are at best an imprecise measure of model quality, especially for discrete choice data like here. To see more clearly whether the model captures theoretically interesting aspects of the data, we again turn to the

---

Notice that θprobably from the simpler model should be mapped onto θlikely in Table 9 because the latter represents the threshold of the inner expressions likely/probably and thus correspond to the simple expression from the first model/data.

---

18
Figure 18: Mean predicted percentages of expression choices in each observation condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts (see Figure 15).
model’s posterior predictive distributions. Figure 18 shows the model’s posterior predictives for the production part. The plot shows the mean counts expected for hypothetical repetitions of the same experiment given the posterior distribution over parameters. The plot also shows the 95% HDIs of these expectations (error bars), together with the empirically observed choice counts (red crosses). Generally the model’s posterior predictive supports the basic observations we made for the production data. As we go from left to right, we see that the model indeed predicts that the conditions in the two leftmost columns mostly trigger expressions ending with *unlikely*; the conditions in the middle column mostly trigger expressions ending with *possible*; and the two rightmost columns show a dominant ending in *unlikely*, at least when there is low higher-order uncertainty. Interestingly, the model’s posterior predictive also supports the observation that *certainly* might be something like an intensifier. For example, just as in the observed data, the model predicts that *certainly likely* is less frequent in condition 6/8 than in 8/8, whereas *is likely* is more frequent in 6/8 than in 8/8; and similarly for *certainly unlikely*.

On the other hand, there are also very clear discrepancies between the data and the posterior predictive, suggesting that the model misses to capture some aspects of the production data. To begin with, in many conditions where the observed choice data has one or two expressions that are selected at much higher rates than the others, the model’s predictions are much less prejudiced. Moreover, the model appears to consistently underpredict the choice rates of *might be possible*, which is far more likely (in the conditions where appropriate) in the empirical data than in the posterior predictive distribution. The expression *might be possible* is logically very weak, which explains why the pragmatic model with its built-in preference for logically stronger expressions, assigns low choice rates to it. Consequently, it is an interesting and puzzling observation that participants seem to like *might be possible* much more than the model expects. We will come back to this observation in the final discussion.

Turning to the interpretation data, the model’s posterior predictives regarding the state, access and observation interpretations are shown in Figure 19. Judging from visual inspection, the model seems to capture the general shape of the empirically observed counts —plotted as red crosses in Figure 19— quite well. This is particularly noticeable for the observation interpretations which have a more distinct shape than the other two dimensions of interpretation. On the other hand, we also clearly see one systematic deviation of the model’s predictions from the empirically observed data. The human data has a much higher choice rate for the option 5 in several conditions of state and access interpretation. For example,
Figure 19: Means of the predicted counts of state, access and observation value choices in each expression condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts (see Figure 16).
with the only exception of probably possible, all expressions containing possible receive an unexpectedly high choice rate for option 5 in both state and access interpretation. This could be an effect of salience of the option 5 which might be perceived as the least committing default option on the scale.

Figure 20 zooms in on the model’s posterior predictions of mean values for state and access interpretation. The observed means, also plotted in Figure 17, are shown as red crosses in Figure 20. The model seems to capture the tendency towards an intensifying effect on the state interpretation of certainly and the down-toning effect of might be in conjunction with likely and unlikely.

On the other hand, the model clearly fails to predict the observed access interpretation of might be possible. According to the observed data, the speaker is taken to be roughly as informed after hearing might be possible as after hearing is possible. The model, in contrast, predicts a too high degree of speaker knowledge (access) for might be possible.

6. Conclusion

We explored the hypothesis that nested uncertainty expressions can be modeled as having a straightforward compositional meaning. Simple expressions express first-order probabilistic information, while nested probability expressions express higher-order uncertainty, namely probabilistic beliefs about probabilistic beliefs. Based on such a compositional semantics, we presented a conservative
extension of a Rational Speech Act model of cooperative communication geared towards maximizing information flow, building on seminal ideas of Grice (1975).

The contribution of this paper is threefold. First, we introduced an experimental scenario, based on draws from an urn, which allowed us to manipulate different levels of uncertainty in a way that is flexible, precise and easy to communicate also visually to participants. We scrutinized the design in Section 3 and concluded that, in due approximation, it is viable to assume that participants collectively behaved in a way which is coherent with a rationalistic normative model of belief formation. This could be interesting on its own and worthy of further investigation, as human reasoners are generally taken to often perform poorly in experimental tasks involving quantities, frequencies and probabilities (e.g., Tversky and Kahneman, 1974; Gigerenzer and Goldstein, 1996; Jones and Love, 2011; Sanborn and Chater, 2016); yet, as the purpose of this paper lies elsewhere, we interpreted this result merely as absence of a clear reason not to adopt a Bayesian model of belief formation into our model of pragmatic language use.

Second, we collected experimental data about English speakers’ use and interpretation of both simple and complex probability expressions in situations of higher-order uncertainty. To the best of our knowledge, this is the first empirical investigation of this kind. The results vindicate our intuition that higher-order levels of uncertainty do influence speakers’ production of probability expressions and that listeners interpret these expressions accordingly. As for simple probability expressions, a noteworthy result is that, given the concrete urn-based context of conversation, participants were able to draw systematic inferences also about the speaker’s mental state of higher-order uncertainty. As for complex probability expressions, our data show that, in the given experimental context, nested expressions are frequently and systematically used to communicate higher-order uncertain information and that listeners are able to draw inferences about the speaker’s higher-order uncertainty also from complex expressions.

Third, we compared the (posterior) predictions of our computational model to the empirical data. The behavior of our model, though not flawless, is promising and insightful. Despite some local discrepancies highlighted by posterior predictive checks, the model captures relevant aspects of the human data fairly well. We take this to mean that our two main conservative modeling assumptions, namely a compositional semantics and Gricean cooperative language use, are at least not outright refuted by the data, but rather help explain general patterns of choice preferences in both production and interpretation tasks. For example, these two assumptions help explain observed speakers’ preferences for complex expressions ending with unlikely, possibly or likely based on their beliefs about the number of...
red balls in the urn, and their preferred choice of a modifying outer expression to attenuate for the second-order probabilistic uncertainty about the number of red balls. These results therefore provide some first empirical evidence in favor of a systematic analysis of nested probability expressions in line with the recent philosophical literature on the topic (e.g., Moss, 2015). Human language users indeed seem to be able to reason systematically about the compositional meaning of complex probability expressions and draw inferences about higher-order uncertainty, in line with the predictions of a normative model of reasoning.

On the other hand, there is also reason to be skeptical of the assumption that any nesting of probability expressions must necessarily receive a straightforward compositional interpretation of the kind we assumed here. The previous sentence contains an instance of a nested modal construction, namely must necessarily, for which a nested modal reading is rather unlikely: must necessarily receive seems near enough synonymous to simple must receive or receives necessarily in this context. This phenomenon, called modal concord, according to which certain double modal constructions reduce to the meaning of a single modal, has attracted some attention in formal semantics (Geurts and Huitink, 2006; Zeijlstra, 2007; Anand and Brasoveanu, 2009). According to the linguistic literature on this topic, modal concord readings can arise in cases where the two modal expressions are similar in logical strength (and in logical type, e.g., both relating to the epistemic state of the speaker or both relating to norms and rules). In the context of our Experiment 3 on complex expressions, there are two candidates for such a potential modal concord reading, namely probably likely and might be possible. Indeed, we saw that is possible and might be possible behaved rather similar in human data, while the model, which assumes a straightforward compositional analysis, had trouble explaining the frequent choice of might be possible in speaker production. The model also had noticeable trouble with the interpretation of might be possible, specifically along the access dimension. It may therefore be an interesting question for further research to what extent such concord readings also apply to nested probability expressions and, moreover, to consider extending the model to also include both a compositional and a modal concord reading as possible interpretations of at least some complex expressions.

Another critical question to ask is whether language users routinely engage in reasoning about higher-order uncertain belief states when they interpret simple or complex probability expressions. The model presented here appears to suggest so, but we do not want to commit to this exaggerated view. Our data suggests that higher-order uncertainty matters to production and interpretation of (complex) uncertainty expressions in a perspicuous but also possibly contrived urn-based sce-
nario. The experimental design made reasoning about higher-order uncertainty relevant (and perhaps easier) in order to address the question whether, in principle and under favorable conditions, language users can reason rationally about higher levels of uncertainty. This an important first step. But now we should investigate further. How frequent or natural is this? What are the circumstances or conditions that make such complex epistemic reasoning relevant to compute?

These questions lead, we believe, to the most pressing open issue which should be addressed in future empirical research. The recent theoretical literature on the interpretation of probability expressions has stressed the importance of looking at the question under discussion (QUD) which an utterance with a probability expression is supposed to address (e.g., Lassiter, 2011; Herbstritt, 2015; Beddor and Egan, 2018). The expression It might be possible that X could either be meant to address an implicit question What is the probability of X? or it could be used to address What do you believe about the probability of X? instead. A listener might have to reason also about which of these (and several other potential QUDs) the speaker may have liked to address (e.g., Kao et al., 2014). Moreover, in order to conclude that a particular utterance is a better fit for one QUD than another, the listener must be able to conceptualize what a good answer to any particular QUD is. In the case of (nested) probability expressions and QUDs like the ones above, this means that listeners should be able to come up with a way of partitioning the set of all possible worlds into distinctions that are relevant in the current context of conversation (e.g., Moss, 2015). Our urn-based scenario gave us a way of fixing these partitions, which is why we are able to state the meaning of expressions in terms of a set of states or a set of observation-access pairs. In normal conversation, this scaffolding based on which to construe possible interpretations for (complex) probability expressions may not at all be clear. In sum, the present paper supports the conclusion that language users can reason systematically and normatively compelling with complex probability and higher-order uncertainty if we fix the interpretation of the relevant context (with the urn-based scenario). The next important question is: how do language users coordinate more flexibly on this contextual scaffolding in which to interpret complex probability expressions?
Appendix A. Population-level rational beliefs

The data-generating model introduced in Section 3 is displayed in Figure A.21 as a probabilistic graphical model, following the conventions outlined by Lee and Wagenmakers (2014).

Figure A.21: The data-generating model as a probabilistic graphical model (left) together with the full formal specification of the model (right). White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence. Boxes indicate scope of indices.

The goal of the model is to test whether we can reasonably assume that the slider ratings observed in the experiment reported in Section 3 could have been produced by individuals holding the normatively correct Bayesian belief about the contents of the urn. The latter is a discrete distribution over the set $S$ of states, i.e., possible quantities $\{0, \ldots, 10\}$ of red balls in the urn and it is defined as follows:

\[
\text{rat.bel}(s|o,a) \propto \text{Hypergeometric}(o|a,s,10) \cdot \text{prior}(s) \quad (A.1)
\]

\[
\kappa - 2 \sim \text{Gamma}(0.01,0.01) \\
\omega \sim \mathcal{U}(0,1) \\
\alpha = \omega \cdot (k-2) + 1 \\
\beta = (1-\omega) \cdot (k-2) + 1 \\
\text{prior}(s) = \text{Betabinom}(s|\alpha,\beta,10) \\
\text{rat.bel}(s|o,a) \sim \text{Hypergeom}(o|a,s,10) \cdot \text{prior}(s) \\
\sigma \sim \text{Gamma}(0.5,1) \\
k \sim \text{Gamma}(5,5) \\
\logit(r_{ivs}) \sim \text{Norm}(\logit(\text{rat.bel}(s|o,a),k),\sigma)
\]
The distribution is a function of partial observations \( v = \langle o, a \rangle \), expressing how likely it is that there are \( s \) red balls in the urn given that the agent has drawn \( a \) balls and observed that \( o \leq a \) among these are red. Equation A.1 is a straightforward application of Bayes’ rule to the Hypergeometric model of the urn, with the prior on \( S \) defined for convenience as a beta-binomial distribution:

\[
\text{prior}(s) = \text{Betabinomial}(s; \alpha, \beta, 10)
\]  

(A.2)

The parametrization of the beta-binomial in terms of \( \alpha = \omega \cdot (\kappa - 2) + 1 \) and \( \beta = (1 - \omega) \cdot (\kappa - 2) + 1 \) is taken from Kruschke (2014), whereas the prior structure on the hyperparameters with \( \kappa - 2 \sim \text{Gamma}(0.01, 0.01) \) and \( \omega \sim \text{U}(0, 1) \) reflects our non-committal stance on the prior and enforces a roughly flat distribution.

The slider ratings recorded in the experiment in each condition are modeled as noise-perturbed realization of the corresponding probability mass prescribed by \( \text{rat.bel} \) in that condition. For each condition, the model predicts a likelihood of observing a particular slider value. In more detail, the likelihood of observing the rating \( r_{ivs} \) given by participant \( i \) in condition \( v = \langle o, a \rangle \) for slider \( s \) is defined as follows:

\[
\text{logit}(r_{ivs}) \sim \text{Norm}(\text{logit}(\text{rat.bel}(s|o,a), k), \sigma)
\]  

(A.3)

Both \( \text{logit}(r_{ivs}) \) and \( \text{rat.bel}(s|v) \) are mapped from the unit interval to the reals, and the observed value is modeled as a realization of the predicted value with normally distributed noise with standard deviation \( \sigma \). The parameter \( k \) modulates the steepness of the logit transform of \( \text{rat.bel}(s|v) \). The prior structure on the parameters \( \sigma \) and \( k \) is borrowed from (Franke et al., 2016).


