Do I know what I want to say? Modeling meaning uncertainty in RSA

Anzi Wang[*1,2], **Carolyn Jane Anderson**[3], **Grusha Prasad**[*2]

[1]Linguistics, Northwestern University
[2]Computer Science, Colgate University
[3]Computer Science, Wellesley College

Correspondence: anziw@u.northwestern.edu, carolyn.anderson@wellesley.edu, gprasad@colgate.edu

Abstract

Models using the Rational Speech Act (RSA) framework typically assume that speakers are certain about the meaning being communicated. In this work we note that there are contexts in which this assumption need not hold, and propose a method (um-RSA) to incorporate this meaning uncertainty within the RSA framework. As a case study, we explore two sources of meaning uncertainty: Counting-Uncertainty (from numerical cognition) and Discounting-Uncertainty (from behavioral economics). We generate predictions from these two hypotheses and test these predictions with two human experiments. The results show that um-RSA can account for differences in uncertainty expression usage that the standard RSA framework cannot account for, thus demonstrating the usefulness of modeling meaning uncertainty.

1 Introduction

The Rational Speech Act (RSA) framework (Frank and Goodman, 2012) views communication as recursive reasoning, where speakers and listeners reason about each others' interpretation of utterances. Prior work has used this recursive reasoning process to effectively model the uncertainty that listeners maintain about speakers and vice versa (Goodman and Stuhlmüller, 2013). Prior work has also modeled the uncertainty that listeners might maintain about properties of the world that is being discussed (Lassiter and Goodman, 2013; Degen et al., 2015; Herbstritt and Franke, 2019) or the question under discussion (Kao et al., 2014a; Kao and Goodman, 2015; Kao et al., 2014b). See Degen (2023) for a review. Despite the differences in the phenomena being modeled, this body of work consistently makes the following assumption: the speaker at the first level of recursion is certain about the target of communication (*target meaning*).

However, this assumption needs not always hold, especially in experiments where participants are presented with images that they are asked to describe. In these experiments, the target meaning for a given trial is often assumed to be specified by the target image presented in that trial, and is thus consistent across speakers (i.e., participants). For example, consider an experiment where participants are presented with images of gumball machines and asked to asked to communicate the probability of the machine dispensing orange gumballs. In a trial where the target image has 18 orange gumballs out of 30, the target meaning is assumed to be $P(red) = 60\%$ (cf. Schuster and Degen 2020).

Yet, uncertainty can emerge in this target meaning based on how the participant *perceives* the target image. Since humans often use an approximate number system (e.g., Gilmore et al. 2011), in a trial with 18 orange gumballs, a participant might approximate the number of red gumballs (e.g., between 15-20), instead of counting to determine the exact number. Such an approximation could result in the participant maintaining uncertainty over the target meaning (e.g., $50\% < P(orange) < 67\%$).

Similarly, uncertainty can also emerge based on how a participant *interprets* the target image. Since people tend to discount low-probability events with high-impact consequences (Sundh, 2024), when presented with such potentially high-impact scenarios, such as election predictions, participants might discount these predictions and assign some mass to lower probabilities. Such discounting, could result in participant maintaining uncertainty over the target meaning (e.g., $P(\text{X winning}) \leq 30\%$).

Drawing on these scenarios, we introduce an extension to the RSA framework that models speakers' uncertainty over the target of communication. We call this extension the *Uncertain Meaning-RSA* or um-RSA.[1] Our approach can be used to model

[*] Equal contribution to project conceptualization and development. AW led the implementation of the RSA models and execution of the human and model experiments. GP led the statistical analyses and writing.

any hypothesis about uncertainty over meaning.

We use um-RSA to generate predictions for the two (not mutually exclusive) hypotheses about uncertainty described above: First, meaning uncertainty can emerge from imprecision in speakers' perception of exact numbers in images (Counting-Uncertainty). Second, meaning uncertainty can emerge from speakers discounting lower probability events (Discounting-Uncertainty). Specifically, we generate predictions for three types of utterances: *bare not*, *might*, and *probably*. The predicted patterns indicate that compared to Counting-Uncertainty, Discounting-Uncertainty results in increased preference for *bare not* and *probably* (i.e., utterances that express more confidence about the outcome), a difference not predicted by the standard RSA framework.

We test these predictions in two within-subject experiments.[2] In the first experiment participants reasoned about images with gumballs and election predictions. We hypothesized participants would discount probabilities in the higher-impact election condition, but not lower-impact gumball condition (cf. Sundh 2024). Consistent with this, we found that participants produced the more confident *probably* and *bare not* utterances more frequently in the election compared to the gumball condition. To further investigate the source of this difference, in a follow-up experiment, we replaced images with text descriptions, thus eliminating uncertainty due to numerical approximations. We found the same qualitative pattern of results, lending support to the hypothesis that the difference between the conditions was likely driven by participants discounting lower probabilities in election predictions.

Thus, our work makes the following contributions. First, we note that the standard RSA framework assumes that speakers are always certain about the target of communication, but this assumption needs not always hold. Next, we propose an extension, um-RSA, that can be used to implement *any* hypothesis about speakers' uncertainty over target meaning. Finally, we demonstrate the usefulness of this approach by generating and testing predictions from two hypotheses.

[1]The 'um' is not just acronym but also captures the sound speakers might make when they are uncertain!

[2]Our model implementation, data, and statistical analyses can be found here: https://github.com/anziw/uncertainty-exp

2 Background

2.1 Rational Speech Act (RSA) Framework

RSA frames conversation as a process in which speakers choose an utterance u given a meaning m they wish to convey, and listeners infer a meaning m given the utterance u they have heard. As a rational framework, RSA assumes that the speaker selects an utterance that maximizes the probability that the listener correctly infers the intended meaning. Accordingly, the framework is built around two core sets: a set of possible meanings a speaker may intend to communicate (M) and a set of possible utterances that can be used to communicate those meanings (U). In many experimental settings, both communicative goals and available utterances are constrained by the design of the task; as a result, RSA models in these contexts typically assume that M and U are finite and relatively small.

Simulating a speaker's probability distribution over utterances requires specifying four components. First, a literal meaning mapping from U to M: for every m_i and utterance u_j, $[\![u_j]\!](m_i)$ is True if the speaker *could* veridically use u_i to convey m_i, and False otherwise. Second, a prior $P(m)$ which captures a listener's subjective probabilities for each meaning in M being the target of communication in the current context. Third, an utterance cost $C(u)$ which specifies the cost of each utterance in U. Finally, a rationality parameter λ which governs how likely the speaker is to choose the utterance that maximizes utility.

As discussed earlier, this framework is also inherently recursive: the speaker reasons about the listener, who in turn reasons about the speaker, who reasons again about the listener, and so on. In this work, we focus on only the first level of recursion, in which a pragmatic speaker S_1 reasons about a literal listener L_0, but the mechanism we propose can be as well extended to higher levels of recursion (see Section 7.1). Simulating S_1 involves two steps. First, the L_0 is simulated by multiplying the literal meaning $[\![u]\!](m)$ by the prior $P(m)$, and normalizing over all meanings to get the probability distribution over meanings conditioned on utterances ($P(m \mid u)$). Second, S_1 is simulated by incorporating the utterance cost $C(u)$ and rationality parameter λ into the L_0 conditional distribution, and then normalizing over all utterances to obtain the probability distribution over utterances conditioned on meanings ($P(u \mid m)$). The specific formulae for these steps are outlined in Table 1.

2.2 Modeling uncertainty with RSA

Prior work studying uncertainty expressions, such as *might* and *probably*, found that there is considerable inter-speaker variability in uncertainty expression usage (Wallsten et al., 1986; Yildirim et al., 2016), and that listeners adapt to these speaker-specific uses (Schuster and Degen, 2020). Schuster and Degen used the RSA framework to model this adaptation by assuming that speakers use different thresholds to determine the literal meaning of these utterances. For example, a confident speaker might consider *probably* to be literally true when describing any event with a probability greater than 60%, whereas a more cautious speaker, adopting a higher threshold, may consider *probably* to be literally true only for events with greater than 80% probability. Listener adaptation to this was modeled as listeners learning these speaker-specific thresholds.

Prior work has also demonstrated the limitations of the RSA framework and proposed extensions. For example, Degen et al. (2015) show that RSA models overestimate the influence of prior world knowledge. Consider the following sentence "Some of the marbles sank", which elicits the scalar implicature that some, but not all of the marbles sank. Standard RSA does not predict this implicature because the prior for all of the marbles sinking is very high given world knowledge. Degen et al. propose an extension in which listeners additionally reason about whether the world being described is normal or "wonky"; if it is a "wonky" world, then they back off to a uniform prior.

As another example, Kao et al. (2014b) show that standard RSA cannot model listeners' imprecise interpretations of non-literal utterances. For example, consider the following utterances:

(1) It took me 62 minutes to get here
(2) It took me 60 minutes to get here
(3) It took me 2 million years to get here

Kao et al. note that listeners only interpret the first utterance literally. Listeners might interpret (2) as the speaker took around, but not exactly, 60 minutes, and (3) as it took the speaker a very long time, but not literally 2 million years. Standard RSA prioritizes utterances that are literally true, and thus cannot model this phenomenon. To overcome this, Kao et al. propose an extension in which speakers can convey affect through their utterances, and can choose to convey the affect or not depending on their goals. In addition to inferring the meaning, listeners also infer speakers' goals and affects.

Despite the different types of variability and uncertainty modeled in prior work and the extensions proposed to the standard RSA framework, there is one central assumption that this body of work makes: the speaker at the first level of recursion is certain about the meaning they intend to communicate. In the following section we discuss why this assumption need not always hold.

2.3 Motivating uncertainty in meaning

There are at least two reasons why a speaker might be uncertain about the meaning being communicated. First, the speaker may be uncertain about the **communicative goal** itself, such as when attempting to express an idea or emotional state that has not yet been fully fleshed out. Second, the speaker may be uncertain about **the information used to form that communicative goal**, such as a speaker who is sure that their goal is to communicate the likelihood of rain based on a weather forecast, but is unsure about how trustworthy the prediction is. In this work, we focus on the latter type of uncertainty. Specifically, we examine two forms of uncertainty that could arise in experimental settings in which participants' intended meanings in any given trial are determined by how they perceive or interpret probabilistic information they are presented with.

Counting-Uncertainty This type of uncertainty could arise when participants determine the intended meaning by estimating counts of objects (e.g.,gumballs) in target images. Prior work on people's numerical representations has demonstrated that for sufficiently large numbers, instead of counting objects individually, people rely on an Approximate Number System (ANS) (Gilmore et al., 2011). Since this system is *approximate*, participants might maintain uncertainty about the exact number of objects in the target image, and thus the exact probability they intend to communicate.

Discounting-Uncertainty This type of uncertainty could arise when participants determine the intended meaning by estimating probabilities of high-stake events (e.g., election outcomes) by reasoning about probabilistic information, such as election predictions. Previous work in behavioral economics and decision making has shown that people have difficulty reasoning about low probability events (Kahneman and Tversky, 1979), and tend to underweight or discount low probabilities for high-impact events (Sundh, 2024). Due to such discounting, participants might maintain uncertainty

Phase	Step	Notation/Equation	Implementation details
Init. world	Define set of observed meanings	M^O	$M^O \in \{m^O_{10\%}, m^O_{20\%} \dots m^O_{90\%}\}$
	Define the set of utterances	U	$U \in \{u_{probably}, u_{might}, u_{not}\}$
	Define set of intended meanings	M^I	$M^I \in \{m^I_{10\%}, m^I_{20\%} \dots m^I_{90\%}\}$
	Initialize weight matrix	$W^{O\to I} \in \mathbb{R}^{\lvert M^O\rvert \times \lvert M^I\rvert}$	$W^{O\to I} \in \mathbb{R}^{9\times 9}$
Init. speaker	Specify literal meaning	$[\![\text{u}]\!](m)$	Sample thresholds $t_{not}, t_{might}, t_{prob}$ $[\![u_{not}]\!](m)$ = 1 if $m \le t_{not}$, else 0 $[\![u_{might}]\!](m)$ = 1 if $m \ge t_{might}$, else 0 $[\![u_{prob}]\!](m)$ = 1 if $m \ge t_{prob}$, else 0
	Specify prior for each meaning	$P(m^O)$	Uniform prior. $P(m^O_i) = \frac{1}{9}$ for all states.
	Specify utterance cost	$C(u)$	Constant cost. $C(u_i) = 3.03$
	Specify rationality parameter	λ	Constant across speakers. $\lambda = 2.211$
	Set weights of $W^{O\to I}$	$W^{O\to I} = W^h$	3 hypotheses (h) to set weights (see § 4.1)
Simulate Committed Speaker	Literal Listener	$L^O_0(m^O \mid u) = norm([\![\text{u}]\!](\text{m}^O) \cdot P(m^O))$	$norm$ normalizes across states
	Move to intended meaning	$L^I_0(m^I \mid u) = (norm(W^{O\to I}) \cdot (L^O_0)^\top)^\top$	$\top$ is transpose; $norm$ across states
	Pragmatic reasoning	$S^I_1(u \mid m^I) = norm(exp(\lambda(logL^I_0(m^I \mid u) - C(u))))$	$norm$ normalizes across utterances
Simulate Uncommitted Speaker	Literal Listener	$L^O_0(m^O \mid u) = norm([\![\text{u}]\!](\text{m}^O) \cdot P(m^O))$	$norm$ normalizes across states
	Pragmatic reasoning	$S^O_1(u \mid m^O) = norm(exp(\lambda(logL^O_0(m^O \mid u) - C(u))))$	$norm$ normalizes across utterances
	Move to intended meaning	$S^I_1(m^I \mid u) = (norm(W^{O\to I}) \cdot (S^O_1)^\top)^\top$	$\top$ is transpose; $norm$ across utterances

Table 1: Formulae for the standard RSA framework (black) and um-RSA (red).

about the exact probability of a high-stakes event occurring given some target prediction, and thus the exact probability they intend to communicate.

Between-participant differences in intended meaning In addition to the *within* participant uncertainty described above, the same observed meaning could result in different intended meanings *between* participants given that prior work has found individual differences in how precisely participants approximate counts (Halberda et al., 2008) or discount low probability events (Khaw et al., 2021)

3 The Uncertain-Meaning RSA: um-RSA

We introduce meaning uncertainty in our model um-RSA by differentiating between two types of meaning: observed meaning (m^O) and intended meaning (m^I). For any given trial t, m^O_t refers to the meaning specified by the information presented in t. Standard RSA models assume that the target of communication for t is m^O_t. In contrast, um-RSA assumes that the target of communication is m^I_t, i.e., the meaning that participants *derive* from the information presented to them in t. Meaning uncertainty emerges from the derivation process.

Specifically, um-RSA defines separate sets for observed and intended meanings, M^O and M^I respectively, and uses a weight matrix $W^{O\to I} \in \mathbb{R}^{\lvert M^O\rvert \times \lvert M^I\rvert}$ to convert probabilities over observed meanings to probabilities over intended meanings. In this work we assume that $\lvert M^O\rvert$ and $\lvert M^I\rvert$ are the same, but this assumption is not necessary.

3.1 Specifying the weights

The weights of $W^{O\to I}$ can be specified in different ways depending on the modeling goal. When the goal is hypothesis testing, the weights can be manually specified to encode specific hypotheses about meaning uncertainty. In this work, we construct three versions of $W^{O\to I}$ to implement our two hypotheses, Counting-Uncertainty and Discounting-Uncertainty, and our baseline No-Uncertainty (see § 4.1). In contrast, when the goal is to infer or characterize meaning uncertainty from data, the weights can be learned using standard optimization techniques. Different approaches for learning these weights are briefly discussed in § 7.1.

3.2 Applying the weights

There are two stages at which $W^{O\to I}$ can be applied in the standard RSA framework, which results in two models of pragmatic speakers that incorporate meaning uncertainty in slightly different ways. First, *a committed speaker model*, where the speaker has uncertainty over the intended meaning initially, but settles on a unique intended meaning before applying pragmatic reasoning to select an utterance. Second, *an uncommitted speaker model*, where the speaker does not commit to a specific intended meaning, but rather marginalizes over possible intended meanings after pragmatic reasoning. We describe these two models below in more detail, and include specific formulae in Table 1.

Committed speaker model Like in the standard RSA framework, the model starts with the literal

listener estimating $L_0^O(m^O \mid u)$, a probability distribution over *observed meanings* conditioned on available utterances. The listener then moves into the intended meaning space by multiplying $L_0^O(m^O \mid u)$ by $W^{O \to I}$, yielding $L_0^I(m^I \mid u)$. Finally, the pragmatic speaker S_1 incorporates the utterance cost and rationality parameter to estimate $S_1^I(u \mid m^I)$, a probability distribution over utterances conditioned on intended meanings. This model maintains the standard RSA assumption that the communicative target is fixed during pragmatic reasoning. Because the model transitions to the intended meaning space before estimating S_1, utterances are evaluated conditioned on intended meanings, effectively committing to an intended communicative target during utterance selection.

Uncommitted speaker model Like in the committed speaker model, this model starts with $L_0^O(m^O \mid u)$. Then, instead of moving into the intended space, this model incorporates the cost and rationality parameter to estimate $S_1^O(u \mid m^O)$, a probability distribution over utterances conditioned on the set of *observed meanings*. Finally, the pragmatic speaker moves into the intended space by multiplying $S_1^O(u \mid m^O)$ by $W^{O \to I}$ to estimate $S_1^I(u \mid m^I)$, a probability distribution over utterances conditioned on *intended meanings*. Since the transition to intended meaning space occurs only after pragmatic reasoning, utterances are optimized without reference to a fixed intended meaning, effectively maintaining uncertainty over communicative targets during pragmatic reasoning.

4 Model experiment

As a case study, we model an experimental setting in which speakers describe 9 possible probabilistic events evenly spaced between 10% to 90% using three possible utterances: not, might, and probably. Thus we operate over the following sets:

- $M^O = M^I = \{m_{10\%}, m_{20\%} \ldots m_{90\%}\}$
- $U = \{u_{not}, u_{might}, u_{probably}\}$

We generate predictions from um-RSA using 3 hypotheses about the type of meaning uncertainty.

4.1 Hypotheses → Weights

No-Uncertainty This hypothesis assumes there is no meaning uncertainty in speakers. Thus observed meaning is identical to the intended meaning. The weight matrix encoding this hypothesis is an identity matrix where the values of the principal diagonal are 1, and all other values 0.

Counting-Uncertainty This hypothesis assumes that speakers are uncertain about intended meaning because their counts of objects in target images are imprecise. Specifically, we assume that when the observed meaning is $P^O(event) = x\%$ the weight matrix has non-zero values for $P^I(event) > x - 10\%$ and $P^I(event) < x + 10\%$, with a value of 1 for $P^O(_{\text{event}}) = x\%$ (i.e., the principal diagonal) The non-zero values are determined by a decay parameter d such that intended probabilities further away from x are assigned lower values.

Discounting-Uncertainty This hypothesis assumes that speakers are uncertain about intended meaning because they discount probabilities in target predictions. Specifically, we assume that when the observed meaning is $P^O(event) = x\%$, the weight matrix has non-zero values for $P^I(event) \leq x$ when $x < 50\%$, and $P^I(event) \geq x$ when $x > 50\%$, with a value of 1 when $P^I(event) = x$. As with Counting-Uncertainty an exponential decay d assigns lower values to probabilities further away from x.

4.2 Generating predictions

We use um-RSA to simulate individual speakers (analogous to participants in a human experiment), and compute the overall predictions by aggregating across all of these speakers. Each speaker is determined by two factors: the thresholds that determine the literal meaning for the three utterances and the decay parameter that determines the non-zero values in the Counting-Uncertainty and Discounting-Uncertainty weight matrices. We sample the thresholds using the following distributions from Schuster and Degen (2020): $t_{not} \sim Beta(0.407, 1.219)$, $t_{might} \sim Beta(0.928, 3.086)$, $t_{probably} \sim Beta(2.552, 1.771)$. Additionally, we enforce the following constraint to capture typical assumptions about the restrictiveness of these expressions: $t_{not} \leq t_{might} \leq t_{probably}$. We sample the decay parameter from $Normal(0, 1)$. Given the thresholds and decay parameter for a participant, for each hypothesis, we follow the sequence of steps outlined in Table 1 to generate the probability of the 3 utterances for all 9 meanings.

4.3 Results

In Figure 1, we present the aggregate predictions for simulated speakers under the three hypotheses (Counting-Uncertainty, Discounting-Uncertainty, and No-Uncertainty) for the two speaker models

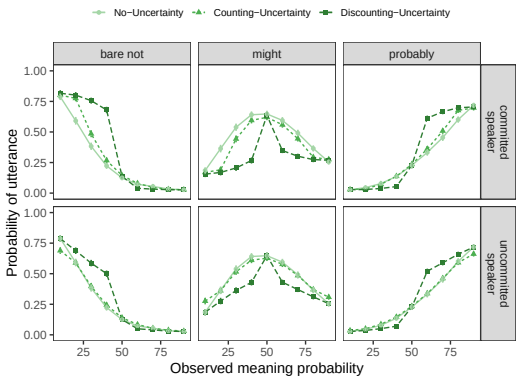

Figure 1: Model predictions under different hypotheses with meaning uncertainty added after the Literal Listener step (top) and after the Pragmatic Speaker step 1. Error bars are two SEs.

(committed speaker and uncommitted speaker). Overall, across the three hypotheses and both speaker models, the simulated speakers displayed patterns consistent with uncertainty expression displayed in Schuster and Degen (2020): $P(not)$ was highest for low probability events, $P(probably)$ was highest for high probability events, and $P(might)$ displayed an inverse u-shape with the highest values for probability events around 50%.

Compared to the No-Uncertainty baseline, the Discounting-Uncertainty condition in both speaker models assigned higher probabilities to utterances which express more confidence towards the event outcome. For events with less than 50% observed probability, Discounting-Uncertainty speakers assigned higher probability to *not* than baseline speakers. Similarly, for events with greater than 50% observed probability, Discounting-Uncertainty speakers assigned higher probabilities to *probably*. The differences between the Discounting-Uncertainty and baseline speakers were greater in the committed speaker model.

In contrast, Counting-Uncertainty and No-Uncertainty condition were considerably more similar to each other, with the conditions being nearly identical in the uncommitted speaker model. While the Counting-Uncertainty condition in the committed speaker model did assign higher higher probabilities to *not* and *probably* for observed probabilities less than and greater than 50% respectively, the qualitative patterns were different from the Discounting-Uncertainty condition.

Taken together, these results predict that when comparing stimuli with the same observed probabilities, stimuli that elicit Discounting-Uncertainty are likely to result in a higher proportion of utterances that express confidence in the outcome (e.g., *not* and *probably*) compared to stimuli that elicit Counting-Uncertainty. This difference is not predicted in the standard RSA framework because it assumes that utterance selection is conditioned on *observed* and not *intended* meanings.

5 Human Experiment 1

The goal of this experiment was to test the predictions from um-RSA. To this extent, we asked participants to reason about gumball machines (cf. Schuster and Degen 2020) and election predictions for an unknown country presented as bars.

Since participants are likely to consider election outcomes to have high impact, and gumball machines to have low impact, we hypothesized that the election, but not the gumball, condition would pattern with Discounting-Uncertainty. Similarly, since probabilistic reasoning about the gumball machine, but not election prediction, requires participants to estimate counts of objects, we hypothesized that the gumball condition, but not the election condition, would pattern with Counting-Uncertainty.

Crucially, since we adopt a within-participant design, for each participant we elicit utterances in the gumball and election conditions for the same observed probabilities. Thus, if participants condition their utterances on *observed probabilities*, like standard RSA assumes, we would not expect to find a difference between these conditions. In contrast, if participants condition their utterances on the derived *intended probabilities*, we would expect the election condition to elicit a higher proportion of *not* and *probably* than the gumball condition.

5.1 Methods

Participants We recruited 120 participants with a US-based IP address, via Prolific. The median completion time was 29 minutes and 47 seconds, and participants received $5.96.

Stimuli and Procedure The experiment was hosted on PCIbex (Zehr and Schwarz, 2018). We used a within-subject blocked design: in the first block participants reasoned about gumball machines, and in the second block about election predictions. Each participant was exposed to 36 trial types: 2 event types (gumball or election), 2 outcome types (for gumball condition, machine dispensing purple or orange gumballs; for election condition, party X or party Y winning), and 9

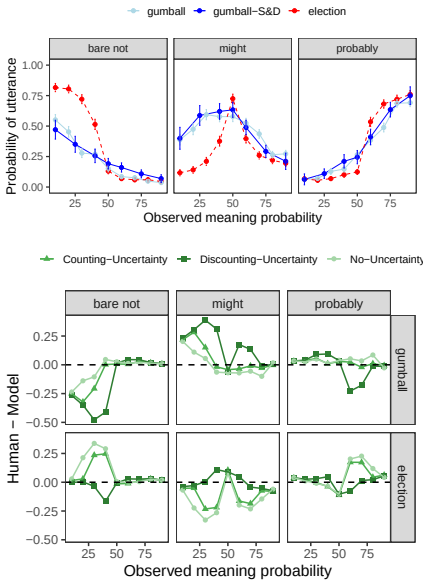

Figure 2: Results from Experiment 1 (top). Average deviation from committed speaker model predictions across 1000 instances (bottom). Error bars are two SEs.

observed probabilities (10% to 90% likelihood of that outcome, in increments of 10%). Participants were presented with each trial type twice, thus resulting in 72 total trials, randomized within each block. Additionally, there were 12 randomly inserted attention checks in each block which asked participants to select the scene they had just seen from two options. At the end of the experiment, participants filled out a demographic survey.

In each trial, participants were presented with a scene, and a person who couldn't see the scene asking a question. In the gumball trials, the scene involved a gumball machine, and the person asked "Will I get a purple/orange gumball?". In the election trials, the scene involved an election prediction, and the person asked "Will party X/Y win the election?". Participants were presented with three possible answers to the questions using *bare not*, *might*, or *probably*. They indicated their relative preference for the utterances with sliders, being required to move at least one slider before proceeding to the next trial. We normalized the slider values into probabilities as a postprocessing step (see § C for a justification). Before each block participants were required to demonstrate that they understood the task, the scene, and how to use the sliders by answering comprehension questions (see § E.2).

Data exclusion We excluded 34 participants because their accuracy on attention checks was below 85% or they reported being non-native speakers.

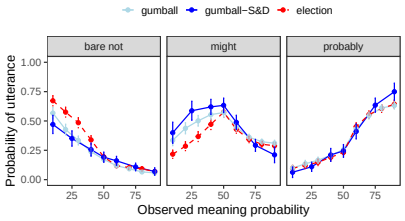

Figure 3: Experiment 2 results with 2*SE error bars.

Statistical analyses Since we were interested in drawing inferences separately for each utterance, instead of using a multinomial logistic regression, we decided to model the probability of each utterance separately with three logistic regression models. Specifically, for each utterance, we fit a mixed effects logistic regression model with the following fixed effects: event type (gumball vs. election), observed probability (50% vs. 10-40%; 50% vs. 60-90%), and the interaction between the two. We included the maximal random effects structure with by-participant random intercept and slopes (see Appendix A).

We decided to model observed probability as a factor with three levels (10-40%, 50%, 60-90%; 50% baseline) instead of a continuous predictor because our experimental design included only 9 observed probabilities, making this predictor not truly continuous. Additionally, for the inferences we want to draw, which are about the overall differences between gumball and election events, the coarse grained distinction of "less than 50%" and "greater than 50%" is sufficient, and thus we did not opt to fit the more complex (and difficult to interpret) mixed effects ordinal logistic regression.

5.2 Results

This experiment uses slightly different stimuli, instructions, and utterance sets from Schuster and Degen (2020) (see § C). Despite these differences, the probabilities of the three utterances in the gumball condition aligned closely (see Figure 2). The slight difference in *bare not* was likely driven by differences in utterance sets (see § C.1). This close replication validates our experimental setup.

Crucially, unlike what standard RSA models predict, there was a significant interaction between event type and observed probability for all three utterances. The effects for *bare not* and *probably* ($P(not \mid 10-40\%) - P(not \mid 50\%)$ and $P(probably \mid 60-90\%) - P(probably \mid 50\%)$) were greater in the election compared to gum-

ball conditions. In contrast, the effects for *might* ($P(might \mid 10-40\%) - P(might \mid 50\%)$ and $P(might \mid 60-90\%) - P(might \mid 50\%)$) was greater in the gumball compared to election condition (see Table 2 for coefficients and p-values).

To compare these empirical results to our um-RSA model predictions, we computed $P_{Human}(u \mid m^O) - P_{Model}(u \mid m^O)$, for each utterance ($u$) and observed meaning ($m^O$) combination, and each of our 3000 model instances (1000 models for each hypothesis). As hypothesized, the Discounting-Uncertainty models better capture the election patterns than the other models (see Figure 2 and § B).

Taken together, these results suggest that speakers are more likely to use expressions that express confidence in the outcome (e.g., *not* and *probably*) when describing election predictions than gumball machines. Our models suggest that this difference is more likely to emerge from the fact that participants *discount* probabilities when reasoning about election predictions. However, since under the committed speaker model, the Counting-Uncertainty condition also resulted in a slight increase in *not* and *probably* compared to the No-Uncertainty condition, we cannot be sure that the difference between gumball and election conditions was not driven by difference degrees of numerical approximation of the observed probability. To test this, we ran a follow up experiment with one crucial difference: we presented participants with text descriptions of the scenes, thus eliminating the need to approximate the observed probability.

6 Human Experiment 2

6.1 Methods

Participants Like with Experiment 1, US-based participants were recruited from Prolific. Each participant received $4.00. The median completion time of the experiment was 27 minutes, 54 seconds.

Stimuli and Procedure We used the same blocked-design and scenes as in Experiment 1, but the images in stimuli were replaced with text descriptions (see § E.3). Initial pilots indicated that these text descriptions significantly increased the duration of the experiment. Therefore, to minimize participant fatigue, we cut down the number of trials by half by asking participants about only one type of outcome, and instead counterbalanced the outcome type between participants (i.e., half the participants were asked questions about orange gumballs and party X, whereas the other half were asked questions about purple gumballs and party Y). The procedure was the same as in Experiment 1.

Data exclusion and Statistical analyses We used the same exclusion criteria and statistical models as in Experiment 1 and excluded 5 participants.

6.2 Results

We see the same qualitative pattern of results as in Experiment 1: for all utterances, we see a significant interaction between observed probability and event type such that *not* is used more frequently in the 10-40%, and *probably* is used more frequently in the 60-90% compared in the election compared to the gumball condition; *might* is used more frequently across the board in the gumball compared to election condition (see Table 3 for coefficients and p-values from our statistical models). Since the text descriptions eliminate Counting-Uncertainty, these results support the inference that the difference between the conditions is likely driven by Discounting-Uncertainty in the election condition.

Note, the preference for *not* and *probably* is smaller in this experiment compared to Experiment 1, especially in the election condition. One possible explanation for this is that discounting of low-probability high-impact events is more likely when the events are experienced rather than described (cf. Sundh 2024), and that our images were more experiential than our text descriptions. This needs to be explored more rigorously in future work.

7 Discussion

In this work, we argue that a common assumption made by most modeling work using the RSA framework — that the speaker is certain about the meaning being communicated — might not always hold: speakers might be uncertain about the meaning (i.e., there is *meaning uncertainty*) either because they are uncertain about the communicative goal itself, or about the information used to form the communicative goal. We propose a generalized approach, um-RSA, where we model meaning uncertainty within the RSA framework by using a weight matrix $W^{O \to I}$ to transform observed meanings to intended meanings. As a case study, we draw on prior work on numerical cognition and behavioral economics to propose two potential sources of meaning uncertainty: Counting-Uncertainty and Discounting-Uncertainty. We use um-RSA to generate predictions from these two hypotheses, and test these predictions with two human experiments.

Our model predictions reveal that for the same observed probabilities, compared to Counting-Uncertainty the preference for utterances that express more confidence about the outcome like *not* and *probably* is higher in Discounting-Uncertainty, a difference not predicted in standard RSA. Consistent with this, our first human experiment revealed that people have a stronger preference for using *not* and *probably* when reasoning about election predictions (hypothesized to elicit Discounting-Uncertainty) than when reasoning about gumball machines (hypothesized to elicit Counting-Uncertainty). Taken together with our modeling results, we tentatively concluded that the difference between the gumball and election conditions was more likely driven by Discounting-Uncertainty than by Counting-Uncertainty. In a follow-up experiment we eliminated Counting-Uncertainty by using text descriptions and found that the difference persisted, thus further supporting our conclusion.

7.1 Future work

Committed vs. Uncommitted speaker models As noted in § 3, there are two different stages at which meaning uncertainty can be incorporated: the committed speaker model and uncommitted speaker model. While these two approaches resulted in qualitatively similar predictions in our experimental setup (see Figure 1), they have different theoretical implications: in the uncommitted model, since the move to intended meaning space occurs *after* pragmatic reasoning, the cost of an utterance can influence the intended meaning. Thus, future work can adjudicate between these two models by experimentally manipulating the cost of utterances.

Other sources of meaning uncertainty Apart from the two sources of meaning uncertainty we considered, there are at least two other categories of meaning uncertainty sources that future work can explore: First, *perceptual uncertainty*, such as from noisy audio or occluded objects in images; Second, *epistemic uncertainty*, such as from confirmation bias or (mis)trust in information sources.

Impact of meaning uncertainty on listeners Since we only focused on the impact of meaning uncertainty on the speaker, this leaves open the question of what impact (if any) this uncertainty might have on listeners. Incorporating meaning uncertainty at the listener level requires first specifying a $W^{O \to I}$ for the listener. There are at least two approaches to specifying this: First, assume that each listener uses their own $W^{O \to I}$, resulting in a potential mismatch between the speaker and listener. Second, assume that listeners adapt to a speaker's $W^{O \to I}$, like they would adapt to other speaker specific preferences. To adjudicate between these approaches, future work can run simulations with and without a mismatch to identify circumstances under which these two approaches generate different predictions, and then test these predictions with human experiments.

Learning the weights of $\mathbf{M^{O \to I}}$ If there is evidence that listeners do adapt to speaker specific $W^{O \to I}$, future work can then explore *what* listeners are adapting to, and *how*. One possibility is to assume a maximally flexible learning mechanism that uses standard optimization techniques to learn each of the weights in the matrix. However, since the number of parameters are directly proportional to the number of utterances and meanings being considered, this approach could prove to be challenging as the spaces of meanings and utterances increase. An alternative is to adopt a more constrained hierarchical approach where a listener first selects the *type* of meaning uncertainty from a small set of hypotheses (e.g., Counting-Uncertainty or Discounting-Uncertainty), and then uses a small set of parameters (e.g., just the decay parameter in our case) to create the $W^{O \to I}$. The learner could then jointly optimize over the hypothesis space, and the parameters for each hypothesis.

7.2 Conclusion

We noted that there are contexts in which speakers are uncertain about the meaning being communicated, and proposed a method (um-RSA) to model this uncertainty within the RSA framework. By using um-RSA to generate predictions about two sources of meaning uncertainty, and testing these predictions with two human experiments, we demonstrated that um-RSA can account for differences in speakers' use of uncertainty expressions that are not predicted by standard RSA models.

Acknowledgments

This work was supported by grants from the Robert H.N. Ho Mind, Brain, and Behavior Initiative at Colgate University. We thank the HSP 2025 audience, Cornell LiMe Lab, Northwestern Experimental Meaning Group and the SEAL Lab, Forrest Davis, Sadhwi Srinivas and Sebastian Schuster for their valuable feedback!

References

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(Volume 9, 2023):519–540.

Judith Degen, Michael Henry Tessler, and Noah D Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Camilla Gilmore, Nina Attridge, and Matthew Inglis. 2011. Measuring the approximate number system. *Quarterly Journal of Experimental Psychology*, 64(11):2099–2109.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

Justin Halberda, Michèle MM Mazzocco, and Lisa Feigenson. 2008. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665–668.

Michele Herbstritt and Michael Franke. 2019. Complex probability expressions and higher-order uncertainty: Compositional semantics, probabilistic pragmatics and experimental data. *Cognition*, 186:50–71.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Justine Kao, Leon Bergen, and Noah Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36.

Justine T. Kao and Noah D. Goodman. 2015. Let's talk (ironically) about the weather: Modeling verbal irony. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 37.

Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Mel W Khaw, Luminita Stevens, and Michael Woodford. 2021. Individual differences in the perception of probability. *PLoS computational biology*, 17(4):e1008871.

Daniel Lassiter and Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of Semantics and Linguistic Theory (SALT)*, volume 23.

Sebastian Schuster and Judith Degen. 2020. I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203:104285.

Joakim Sundh. 2024. Human behavior in the context of low-probability high-impact events. *Humanities and Social Sciences Communications*, 11(1).

Thomas S Wallsten, Samuel Fillenbaum, and James A Cox. 1986. Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, 25(5):571–587.

Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, and T. Florian Jaeger. 2016. Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, 87:128–143.

Jeffrey Zehr and Florian Schwarz. 2018. Penncontroller for ibex. `https://github.com/PennController/pcibex`. Accessed: 2025-04-26.

A Statistical analyses

A.1 Model details

Model formula prob_utterance $\sim$ Meaning * Event type + (1 + Meaning * Event type | randomid)

Coding of factors

- Meaning: Factor with three levels (10-40%, 50%, 60-90%), resulting in two predictors.
  - Meaning 1: 50% coded as 0; 10-40% coded as 1.
  - Meaning 2: 50% coded as 0; 60-90% coded as 1.
- Event type: Factor with two levels (Election coded as 0, Gumball coded as 1).

A.2 Results from statistical models

Utterance	Coefficient	Estimate	p-value
Not	Meaning 1 (50% vs. 10-40%)	2.897	1.35e-09
	Meaning 2 (50% vs. 60-90%)	-1.582	0.153
	Event type (Gumball vs. Election)	-0.921	0.232
	Meaning 1 : Event type	3.272	2.23e-05
	Meaning 2 : Event type	2.008	0.259
Might	Meaning 1 (50% vs. 10-40%)	0.282	0.3194
	Meaning 2 (50% vs. 60-90%)	-1.415	6.25e-09
	Event type (Gumball vs. Election)	0.982	0.0055
	Meaning 1 : Event type	-4.113	< 2e-16
	Meaning 2 : Event type	-1.835	3.70e-06
Probably	Meaning 1 (50% vs. 10-40%)	-1.683	2.09e-08
	Meaning 2 (50% vs. 60-90%)	2.140	< 2e-16
	Event type (Gumball vs. Election)	-8.561	5.92e-08
	Meaning 1 : Event type	8.177	2.71e-07
	Meaning 2 : Event type	9.249	1.01e-08

Table 2: Results from Experiment 1. A separate model was fit to each utterance.

Utterance	Coefficient	Estimate	p-value
Not	Meaning 1 (50% vs. 10-40%)	10.641	< 2e-16
	Meaning 2 (50% vs. 60-90%)	3.806	< 2e-16
	Event type (Gumball vs. Election)	6.944	< 2e-16
	Meaning 1 : Event type	-6.198	< 2e-16
	Meaning 2 : Event type	-4.270	< 2e-16
Might	Meaning 1 (50% vs. 10-40%)	0.531	0.0114
	Meaning 2 (50% vs. 60-90%)	-1.882	3.12e-11
	Event type (Gumball vs. Election)	-0.210	0.4265
	Meaning 1 : Event type	-0.783	< 0.0114
	Meaning 2 : Event type	0.09413	0.7629
Probably	Meaning 1 (50% vs. 10-40%)	-0.535	0.136
	Meaning 2 (50% vs. 60-90%)	2.711	< 2e-16
	Event type (Gumball vs. Election)	-1.040	0.0433
	Meaning 1 : Event type	0.751	0.2152
	Meaning 2 : Event type	1.046	0.047

Table 3: Results from Experiment 2. A separate model was fit to each utterance.

B Additional results of Experiment 1

Figure 4 replicates Figure 2 with predictions from the uncommitted speaker model.

C Differences between Schuster and Degen (2020) and our experiments

Here are the different ways in which our experimental setup differed from Experiment 1 in Schuster and Degen (2020).

The scene In the scene Schuster and Degen used, the gumball machine was on a table. A little girl who couldn't see the machine tells a man who can see the machine "I want an orange one". The interaction between the child and the man was important for Schuster and Degen because they wanted to eventually study how listeners adapt to different speakers (e.g., change the man in the scene to a woman). Since we only wanted to model the speaker, we did not need such a complex setup. In our scene, the gumball machine was on a table. A person (represneted with a stick figure) who was behind a wall and could not see the machine asked "Will I get an orange gumball?".

Utterance sets Schuster and Degen were interested in collecting participants preferences for a wide range of uncertainty expressions. Therefore, they ran several sub-experiments where they varied the specific utterances used. In each sub-experiment, there were always two uncertainty expressions (e.g., might-probably, probably-not, might-not, not-could, might-could, etc), and a third "other" option. In our work, we were interested in comparing *might*, *probably*, and *not*. We decided to not have an "other" option because it was hard to interpret *why* a participant might choose this option.

Meaning set Schuster and Degen used the following meaning set: 0%, 10%, 25%, 40%, 50%, 60%, 75%, 90%, 100%. In our work we wanted to avoid the 100% because we did not have an "other" option, and thus none of our three utterances could be true. For symmetry, we also excluded the 0%. We kept the meaning set size the same by having more fine-grained increments. Thus our meaning set was: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%.

Sliders Schuster and Degen asked participants to assign points to each of the utterances, and the sliders automatically jumped back if participants assigned more than 100 points. Our pilot experiments suggested that this task was too tedious and participants were more likely to pick whole numbers (to make the

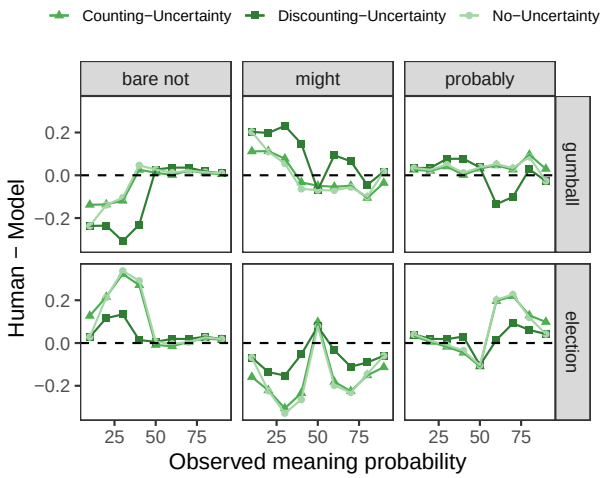

Figure 4: Deviation from uncommitted speaker model predictions averaged over 1000 model instances. Error bars are two SEs.

addition easier). To better capture participants initial instincts, we asked participants to indicate their relative preference and then later normalized these values into probabilities.

C.1 Why was the probability of bare *not* slightly different?

Since Schuster and Degen did not have an experiment where all three of our utterances were present, we selected the *might* and *probably* probabilities from their *might-probably* experiment, and the *bare not* probability from their *might-not* experiment. Since the *might-not* experiment did not have *probably*, the utterance that tends to be most preferred for higher target probabilities, it makes sense that the probability assigned to bare *not* in these cases was higher than in our experiment which did have *probably* as a possible utterance. The close replication in all other cases suggest that the other differences in experimental setup did not impact our results.

D Literal meaning example

	$m_{10\%}$	$m_{20\%}$	$m_{30\%}$	$m_{40\%}$	$m_{50\%}$	$m_{60\%}$	$m_{70\%}$	$m_{80\%}$	$m_{90\%}$
$u_{probably}$	0	0	0	0	0	1	1	1	1
u_{might}	0	0	0	1	1	1	1	1	1
u_{not}	1	0	0	0	0	0	0	0	0

Table 4: Literal meaning map of a hypothetical participant who samples threshold of 10% for not, 30% for might, 60% for probably

E Experimental setup

E.1 Instructions

Instructions in Experiments 1 and 2 were identical except that stimuli were labeled *images* and *scenes*, respectively.

Start of experiment Welcome! In this experiment, you will be presented with images/scenes and be asked to answer questions about each image/scene. For each of A's questions, you will also be presented with three possible responses. Your task is to decide how much you prefer each of the three utterances as a response to A's questions.

Gumball condition In this part of the experiment, you will be presented with images/scenes of gumball machines. The gumball machines are filled with purple and orange gumballs. The gumballs will be tossed around before a random one is dispensed. Here is an example of what a gumball machine may look like.

Election condition In this part of the experiment, you will be presented with images/scenes of election predictions from an unknown country. In this country, the two major parties, Party X and Party Y, compete for votes in different counties. In each county, the party with the maximum votes wins that county. A bipartisan company is interested in studying the probability of the two parties winning in different parts of the country – specifically they are interested in comparing the results in different counties. So they generate predictions about the outcomes of the elections county by county. This company has a great track record of generating very reliable predictions. Here is an example of what a prediction may look like.

Hypothetical listener "A" In this experiment, you will also see a fictional person "A". The fictional person "A", who cannot see the images/scenes, will ask you questions about the images/scenes. Here is an example of what A may look like.

Sliders In the experiment, you will be presented with three possible responses for each of A's questions. For each response, you will use a slider to indicate your degree of preference for the response. Each slider will start at the far left, which indicates zero preference. If you would never use an utterance to answer A's question, leave its slider at the far left. If you would always pick an utterance over the other two utterances, move its slider to the far right and leave the other two sliders in their original positions. If all the sliders are set to the same position, it means you are equally likely to pick any of the utterances to answer A's question.

E.2 Understanding checks

See Figure 5 and 6.

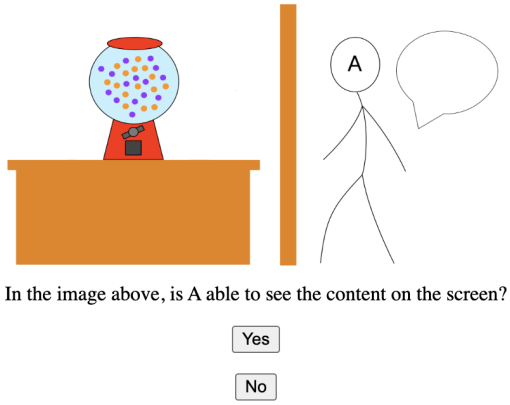

Figure 5: Understanding check for fictional person/ hypothetical listener "A".

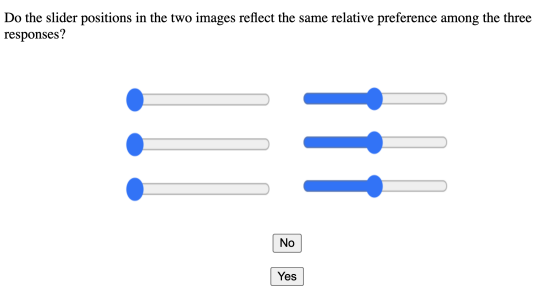

Figure 6: Understanding check for the sliders.

E.3 Example trials

See Figure 7 through 12.

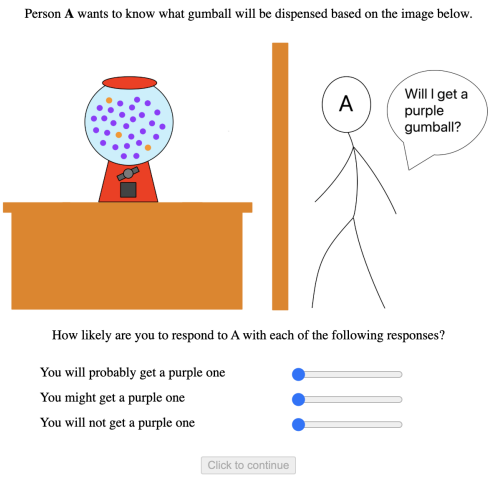

Figure 7: A sample trial of the gumball condition from Experiment 1 (image).

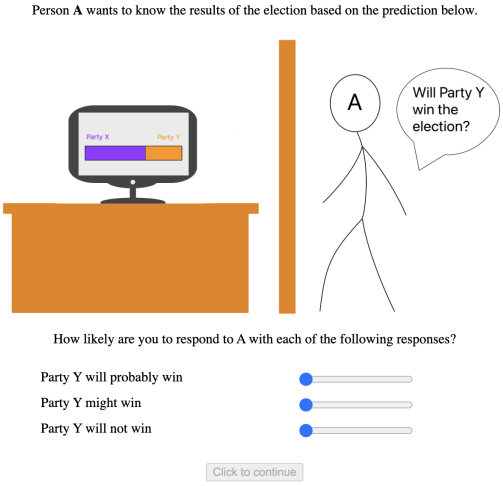

Figure 8: A sample trial of the election condition from Experiment 1 (image).

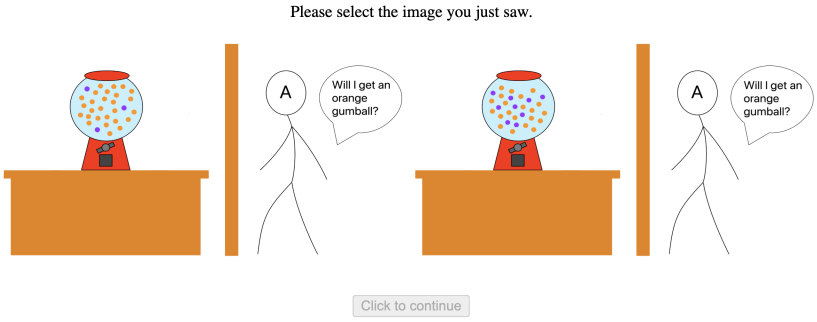

Figure 9: A sample attention check from Experiment 1 (image).

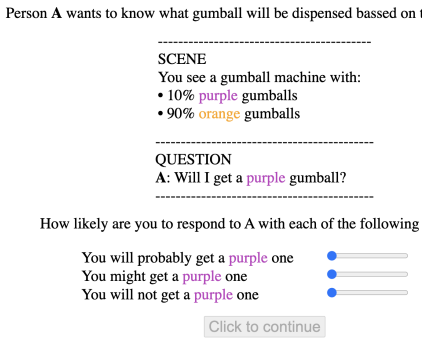

Figure 10: A sample trial of the gumball condition from Experiment 2 (text).

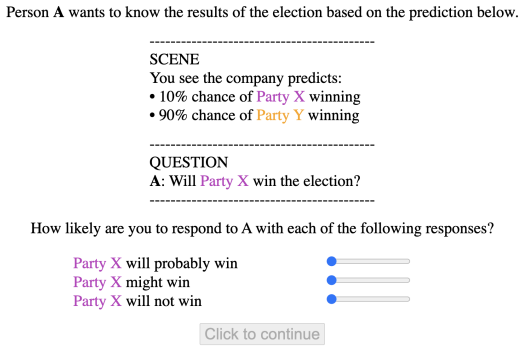

Figure 11: A sample trial of the election condition from Experiment 2 (text).

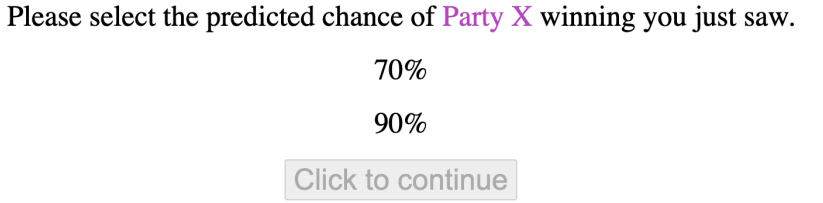

Figure 12: A sample attention check from Experiment 2 (text).