# I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions

Sebastian Schuster and Judith Degen
Department of Linguistics, Stanford University

## Abstract

Pragmatic theories of utterance interpretation share the assumption that listeners reason about alternative utterances that a speaker could have produced, but didn't. For such reasoning to be successful, listeners must have precise expectations about a speaker's production choices. This is at odds with the considerable variability across speakers that exists at all levels of linguistic representation. This tension can be reconciled by listeners adapting to the statistics of individual speakers. While linguistic adaptation is increasingly widely attested, semantic/pragmatic adaptation is underexplored. Moreover, what kind of representations listeners update during semantic/pragmatic adaptation – estimates of the speaker's lexicon, or estimates of the speaker's utterance preferences – remains poorly understood. In this work, we investigate semantic/pragmatic adaptation in the domain of uncertainty expressions like *might* and *probably*. In a series of web-based experiments, we find 1) that listeners vary in their expectations about a generic speaker's use of uncertainty expressions; 2) that listeners rapidly update their expectations about the use of uncertainty expressions after brief exposure to a speaker with a specific usage of uncertainty expressions; and 3) that listeners' interpretations of uncertainty expressions change after being exposed to a specific speaker. We present a novel computational model of semantic/pragmatic adaptation based on Bayesian belief updating and show, through a series of model comparisons, that semantic/pragmatic adaptation is best captured by listeners updating their beliefs both about the speaker's lexicon and their utterance preferences. This work has implications for both semantic theories of uncertainty expressions and psycholinguistic theories of adaptation: it highlights the need for dynamic semantic representations and provides evidence against accounts that cast adaptation as simple low-level priming.

*Keywords:* adaptation; language comprehension; experimental pragmatics; Bayesian cognitive modeling; uncertainty expressions

## 1   Introduction

One of the key assumptions about pragmatic reasoning is that listeners reason about alternative utterances when interpreting a speaker's utterance (Grice, 1975; Horn, 1984). For example, consider the following sentences that give rise to scalar implicatures.

(1)   a.   Alex: Bill ate some of the cookies.

     b.   ⤳ Bill did not eat all of the cookies.

(2)   a.   Tom: The movie was okay.

     b.   ⤳ The movie was not great.

(3)   a.   Sue: It might snow tomorrow.

     b.   ⤳ It is not certain that it will snow tomorrow.

According to Gricean pragmatic theories, listeners assume that a speaker is cooperative and arrive at the inference in (1b) through a counterfactual reasoning process: they reason that if Alex had wanted to communicate that Bill ate all of the cookies, Alex would have uttered the more informative statement *Bill ate all of the cookies.* Assuming that Alex knew the truth regarding the more informative sentence, it must be that the more informative statement is not true, which leads the listener to conclude (1b). Analogous reasoning leads to the inferences in (2b) and (3b).

Accounts of pragmatic reasoning share the implicit assumption that listeners have precise expectations about the speaker's language use – specifically, which utterance alternatives were available to the speaker that they didn't use – in different situations. Listeners can only draw correct pragmatic inferences if they know what a speaker would have said to communicate alternative world states. Arguably, this assumption is valid in many contexts – after all, languages are highly conventional systems (Lewis, 1969). However, language users also exhibit a great deal of variability in their phonetic, lexical, and syntactic choices (e.g., Allen, Miller, & DeSteno, 2003; Finegan & Biber, 2001; Harrington, Palethorpe, & Watson, 2000; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Weiner & Labov, 1983). For instance, the variability in use of quantifiers such as *some* and *many* and uncertainty expressions like *might* and *probably* is reflected in listeners' variable expectations of quantifier use (Yildirim, Degen, Tanenhaus, & Jaeger, 2016) and in considerable inter-subject variability in the interpretation of uncertainty expressions (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). This raises a puzzle: how can we reconcile the assumption of stable utterance alternatives required for capturing pragmatic inferences with what appears to be rampant variability in speakers' actual language use?

Recent work suggests that listeners deal with variability in language use by adapting to it, i.e., by updating their expectations about a speaker's likely productions (e.g., Fine & Jaeger, 2016; Kamide, 2012; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003). In the domain of semantics/pragmatics, this is a process known as *semantic/pragmatic adaptation.* In a series of experiments, Yildirim et al. (2016) exposed participants to different speakers whose use of the quantifiers *some* and *many* varied in descriptions of quantities of candies of a particular color like *Some of the candies are green.* After exposure to a speaker, they probed participants' expectations about the

speakers' likely descriptions of different quantities of green candies and found that participants indeed formed speaker-specific expectations. However, while the results consistently suggest that listeners update *some* type of expectations, the nature of the expectations that listeners update is unknown. In particular, it is an open question whether this kind of semantic/pragmatic adaptation is a result of listeners learning speaker-specific utterance *preferences* or whether listeners form speaker-specific *semantic representations*. Answering this question about the nature of adaptation is the focus of the work reported here.

As a starting point for this investigation, we consider adaptation in other linguistic domains. Apart from the work on quantifiers, linguistic adaptation has been observed in phonetics (Babel, 2012; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Goldinger, 1998; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2005, 2007; Norris et al., 2003), syntax (Fine & Jaeger, 2016; Fine, Jaeger, Farmer, & Qian, 2013; Kamide, 2012; Kroczek & Gunter, 2017; Myslín & Levy, 2016),[1] intonation and prosody (Kurumada, Brown, & Tanenhaus, 2012; Roettger & Franke, 2019), and with phenomena such as referring expressions (Brennan & Clark, 1996; Brennan & Hanna, 2009; Clark & Wilkes-Gibbs, 1986; Hawkins, Frank, & Goodman, 2017; Horton & Gerrig, 2005; Metzing & Brennan, 2003), contrastive inferences (Grodner & Sedivy, 2011; Pogue, Kurumada, & Tanenhaus, 2016), and lexical associations (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019).

For most of these phenomena, there is converging evidence regarding the representations that listeners update during adaptation. At the phonetic level, listeners update (at least) their expectations about speakers' **mapping** between acoustic cues and phonemes (e.g., Kleinschmidt & Jaeger, 2015). At the syntactic level, listeners update (at least) their expectations about speakers' **preferences** for different syntactic structures. In contrast, at the semantic/pragmatic level the adaptation process and the nature of the updated representations is still poorly understood. This is not surprising considering that it is challenging to directly probe beliefs about semantic representations or beliefs about speaker preferences without a model that can quantitatively link behavioral data to these beliefs.

Recent advances in probabilistic modeling of pragmatic language understanding within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016) allow us to formally investigate the two likely candidates for representations that are updated during semantic/pragmatic adaptation mentioned above: *utterance preferences* and *semantic representations*. To elaborate, listeners might update their beliefs about speakers' preferences for producing a particular expression (e.g., a preference for *might* over *probably*) – analogous to syntactic adaptation. Alternatively, listeners might update their beliefs about a speaker's lexicon, i.e. their mapping between words and world states (e.g., the range of event probabilities that *probably* is compatible with) – analogous to phonetic adaptation. Finally, listeners might track both preferences and mappings.

To illustrate how different beliefs about lexica and utterance preferences can lead to different interpretations, consider the interpretation of the uncertainty expression *probably* produced by three different hypothetical speakers. For the sake of this example, let us assume the only three expressions that a speaker can choose from are *might*, *probably*, and

---

[1]Note, however, that some of these studies failed to replicate and it is still unclear under what circumstances syntactic adaptation can be observed (see Harrington Stack, James, & Watson, 2018; Liu, Burchill, Tanenhaus, & Jaeger, 2017).
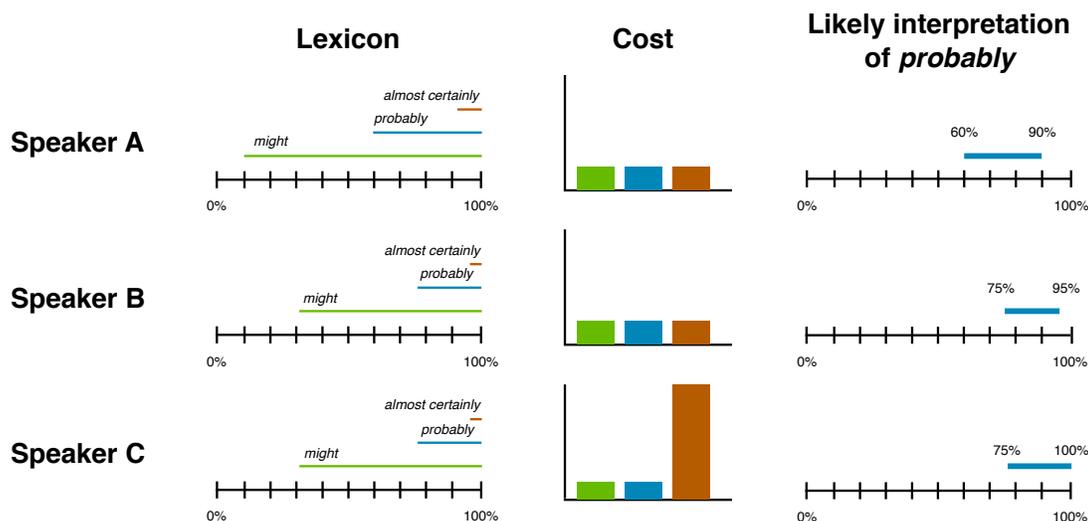
*Figure 1*. Lexica, utterance preferences and likely interpretation of *probably* for three different hypothetical speakers. The region of the probability scale covered by each line in the Lexicon panel indicates the corresponding expression's literal semantics. Height of bars in the Cost panel indicates the speaker's cost (dispreference) for each expression.

*almost certainly.* A listener's beliefs about the three speakers' lexica and preferences are schematically illustrated in Figure 1.

First, consider speaker **A**, for whom *might* is semantically felicitous if the described event probability (e.g., of snowing) exceeds 10%, *probably* if the event probability exceeds 60% and *almost certainly* if the event probability exceeds 90%. If a listener has accurate beliefs about **A**'s mapping between expressions and event probabilities and observes **A** produce the sentence *It will probably snow*, they will be likely to infer a probability of snowing between 60 and 90%. As illustrated above, the reasoning follows the schema of a standard scalar implicature (Grice, 1975; Horn, 1984): if **A** had intended to communicate a probability above 90%, they could have said *It will almost certainly snow*, which would have been more informative and equally relevant. Assuming the speaker knows the actual event probability and is cooperative, it is therefore likely that the intended probability is not above 90%.[2]

Now, consider speaker **B**, for whom *might* is semantically felicitous if the event probability exceeds 30%, *probably* if the event probability exceeds 75% and *almost certainly* if the event probability exceeds 95%. If a listener has accurate beliefs about **B**'s mappings, they will be likely to infer, via the same reasoning as above, a chance of snow between 75% and 95% when they hear **B** produce the same sentence, *It will probably snow*.

Finally, consider speaker **C**. **C** uses the same mapping between expressions and event probabilities as **B**. However, **C** has a strong preference against producing *almost certainly*.

---

[2]Under a standard Gricean view, the negation of the stronger alternative is inferred categorically. However, we adopt probabilistic language here in keeping with recent results that scalar inferences are more aptly viewed as probabilistic inference under uncertainty (Goodman & Stuhlmüller, 2013).

If a listener has accurate beliefs about **C**'s lexicon and production preferences, they will be likely to infer a chance of snow between 75% and 100% when they hear **C** produce *It will probably snow* since they will not consider *almost certainly* a likely alternative. That is, the scalar inference will be blocked by the additional knowledge of the speaker's production preferences.

Thus, a listener who tracks the variability in these hypothetical speakers' lexica and production preferences will draw on average more accurate inferences about the world than one who commits to a particular lexicon without updating it. We investigate the nature of the representations updated during semantic/pragmatic adaptation in the the domain of uncertainty expressions, i.e., words or phrases that can be used to express uncertainty, as used in descriptions of potential future events. These expressions include epistemic modals such as *might*, *probably*, and *could* (see, for example, Hacquard, 2011; Kratzer, 1991) but also phrases such as *it looks like*, which have been primarily investigated in the experimental pragmatics literature (e.g., Kurumada, Brown, Bibyk, Pontillo, and Tanenhaus, 2014; Pogue and Tanenhaus, 2018).

Uncertainty expressions have several properties that make them a good testing ground for studying semantic and pragmatic adaptation. First, there is no consistent mapping between uncertainty expressions and event probabilities (e.g., Clark, 1990; Pepper & Prytulak, 1974), which suggests that listeners have to rely on additional contextual information (such as speaker identity) if they want to infer an event probability that a speaker intended to communicate using an uncertainty expression. Second, there is considerable inter-speaker variability in the use of these expressions (Wallsten et al., 1986) and therefore it is likely that listeners expect different speakers to use these expressions differently. Lastly, interpreting uncertainty expressions plays an important role in many everyday situations from the banal – such as talking about the weather – to the serious – such as communicating about health risks (Berry, 2004; Lipkus, 2007; Politi, Han, & Col, 2007) or making financial decisions (Doupnik & Richter, 2003). Thus, listeners would benefit from tracking how a given speaker uses these expressions.

In order to establish the nature of the representations that are updated during adaptation to variable use of uncertainty expressions we proceed through the following steps:

1. Quantify the variability in listeners' expectations about a generic speaker's production of uncertainty expressions (Experiment 1, Section 2).

2. Propose a probabilistic computational pragmatics model of production expectations about uncertainty expressions that functions as proxy for listeners' baseline generative model of a generic speaker. Evaluate the model on the data from Experiment 1 (Section 3). The model is formulated within the Rational Speech Act framework, a probabilistic formalization of Gricean pragmatic reasoning (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016).

3. Measure whether and to what extent listeners update their expectations when exposed to a speaker who is either more cautious or more confident in their use of uncertainty expressions than the baseline speaker model (Experiment 2, Section 4).

4. Extend the baseline model to support learning in interaction. Create three versions of this model which differ in terms of which model components are updated in response

to exposure to simulated cautious and confident speakers: update of only production preferences, only the lexicon, or both. Use model comparison between these three adaptation models as a hypothesis-testing tool to infer which representations undergo adaptation, by evaluating which adaptation model best captures the observed post-exposure expectation data from Experiment 2 (Section 5).

5. To further test the adaptation models, use them to derive predictions about post-exposure interpretation. Measure interpretation and evaluate the model (Experiment 3, Section 6).

We find that listeners indeed update their beliefs about different speakers' use of uncertainty expressions, and that this adaptation is reflected both in post-exposure measures of production expectations and interpretation. The data are best captured by the adaptation model in which both the lexicon and the speaker's production preferences are updated. We conclude with a discussion of remaining open questions and the implications of our findings for theories of interactive (e.g., Pickering & Garrod, 2004, 2013) and partner-specific language processing (e.g., Horton & Gerrig, 2005, 2016; Metzing & Brennan, 2003).

## 2 Experiment 1: Pre-exposure ratings

We first conducted a norming study, which served the following theoretical and methodological purposes. First, it served as a methodological check on whether the paradigm is suited for manipulating fine-grained event probabilities. Second, it addressed the theoretical question of whether listeners vary in their expectations about a generic speaker's use of uncertainty expressions, by collecting participants' judgments about uncertainty expressions they expected speakers to use for varying probabilities of receiving gumballs of a particular color from a gumball machine. Third, the results from this study informed the experimental design of the adaptation experiments reported in later sections, by allowing us to both choose which pair of uncertainty expressions to test adaptation on, and to determine the particular event probability for which participants had roughly equiprobable expectations about which expression of uncertainty a generic speaker would use to report an event with that probability. Lastly, we used the data collected in this study to estimate population-level prior beliefs for the adaptation model reported in Section 5.

### 2.1 Participants

We recruited a total of 420 participants (20 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $1.80 (condition 1), $1.50 (conditions 2-15), or $2.00 (conditions 16-21), depending on the number of trials, which amounted to an hourly wage of approximately $12–$15.

### 2.2 Materials and Procedure

This study was a forced-choice production experiment. Participants were instructed that over the course of the experiment, they would see several scenes with an adult man, a young girl, and a gumball machine on a table and that the gumball machine is too high up on the table for the girl to see (see Figure 2 for an example scene). After completing an
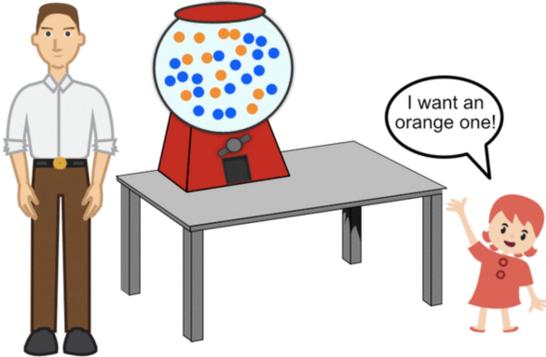
*Figure 2*. Example trial from Experiment 1.

attention check which asked participants whether the girl could see the gumball machine,[3] participants saw a series of scenes and were asked to rate how likely they thought it was that the adult would produce two given responses by distributing 100 points across the two given utterances and the blanket *something else* option (OTHER). Sliders automatically jumped back if participants tried to distribute more than 100 points. In each scene, the child uttered *"I want a blue one"* (target color: blue) or *"I want an orange one"* (target color: orange), randomized across participants.[4] The gumballs in the machines were tossed around continuously to prevent participants from counting the gumballs and to make sure that participants did not consider it more likely to get one of the gumballs at the bottom of the machine. In each of the 21 conditions, participants saw only two of the following seven possible adult utterances with different uncertainty expressions:

- You'll get a blue/orange one. (BARE[5])

---

[3]Participants had to go back to the instructions in case they responded incorrectly. This was the case for 41 participants.

[4]In condition 1 (*bare-might*), as well as conditions 16-21 (all conditions with *bare not*), the target color was randomized across trials. While randomization of the target color across trials increased the correlation between the ratings for the two colors, the average ratings for each condition independent of the target color were not affected by this choice. See Appendix A or a detailed discussion of the effect of this manipulation on the ratings.

[5]As a notational convention, we refer to utterances with uncertainty expressions in SMALL CAPS and to the uncertainty expression itself in *italics*.
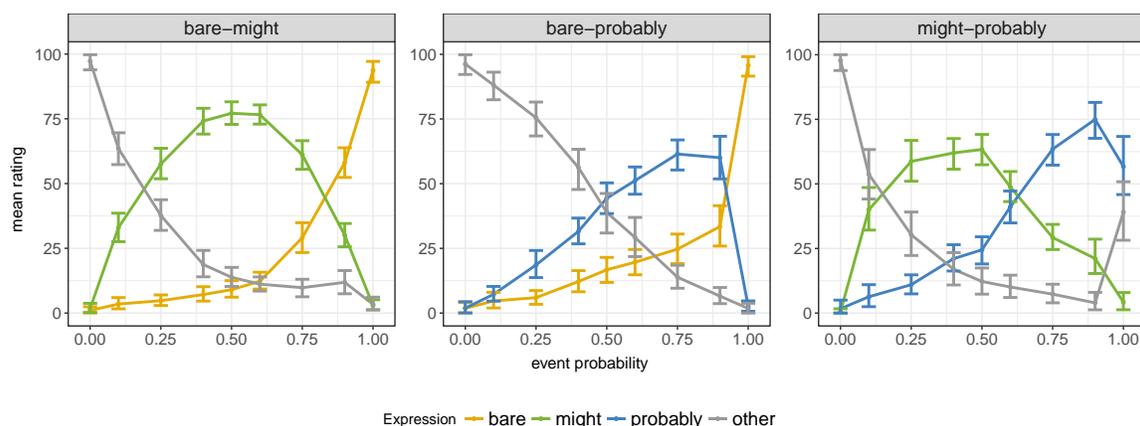
*Figure 3*. Results from 3 conditions of Experiment 1. Error bars correspond to bootstrapped 95%-confidence intervals.

- You might get a blue/orange one. (MIGHT)

- You'll probably get a blue/orange one. (PROBABLY)

- I think you'll get a blue/orange one. (THINK)

- It looks like you'll get a blue/orange one. (LOOKS LIKE)

- You could get a blue/orange one. (COULD)

- You won't get a blue/orange one. (BARE NOT)

Within each condition, we manipulated the percentage of target color gumballs across trials, which we take as proxy for the objective probability of receiving a gumball of the target color. Each participant saw 3 trials[6] for each of the following percentages: 0%, 10%, 25%, 40%, 50%, 60%, 75%, 90%, 100%. We randomized the order of expressions across participants and trials were presented in randomized order.

**2.2.1   Results and Discussion.**    Figure 3 shows participants' ratings for different gumball proportions for 3 of the 21 conditions, namely all combinations of the conditions with the utterances BARE, PROBABLY, and MIGHT (see Appendix B for the results from the other 18 conditions). The results from these three conditions highlight several important properties of participants' behavior in this experiment that generalize to all conditions. First, the ratings for individual utterances are influenced by the utterance choices presented to participants. If we compare the ratings for MIGHT in the *bare-might* and the *might-probably* condition, we see that MIGHT received high ratings for a larger range of event probabilities when it is paired with BARE than when it is paired with PROBABLY. We observe similar effects for the other two utterances. This suggests that participants are cued

---

[6]In condition 1 (*bare-might*), participants saw each gumball machine 6 times – 3 times when being asked to produce a statement about orange gumballs and 3 times when being asked to produce a statement about blue gumballs. In conditions 15-20 (all conditions with *bare not*), participants saw each machine 4 times: 2 times for each color.
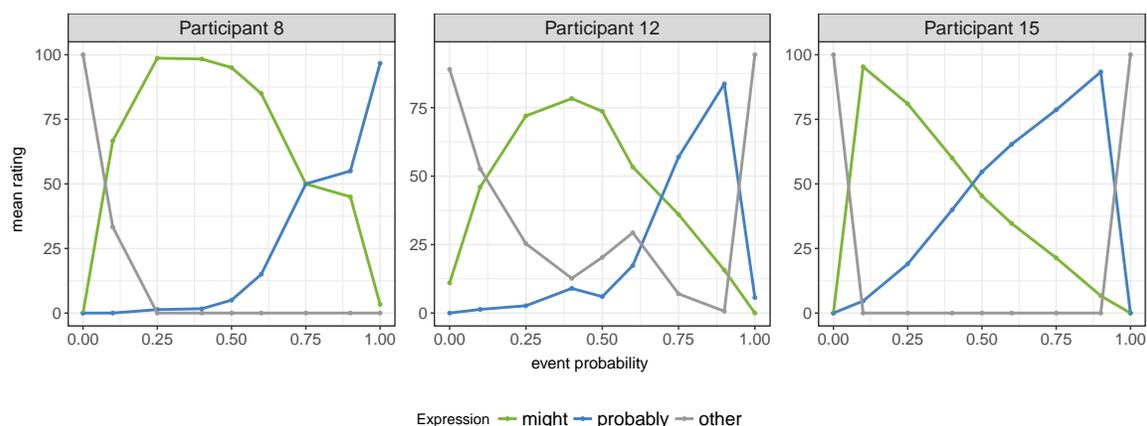
*Figure 4*. Results of three individual participants in the *might-probably* condition of the Experiment 1.

towards using the utterances provided in the experiment and that their ratings depend on the presented alternatives – an effect that has also been observed for quantifiers (Degen & Tanenhaus, 2016).

Second, the results suggest that participants are sensitive to the different event probabilities and that this paradigm is well suited to study the mapping between event probabilities and uncertainty expressions. For example, in the *might-probably* condition, participants provided considerably different ratings when they were presented with a gumball machine with 50% target color gumballs than when they were presented with 60% target color gumballs.

Third, in all conditions, the mean ratings are graded and except for the 0% and 100% target color gumball trials, the average rating for none of the utterances is close to 100. There are two potential explanations for this observation. It could be that participants provided categorical ratings, i.e., generally assigned 100 points to one of the three options but the category boundaries vary across participants which leads to the graded average ratings. It could also be that participants' individual ratings are graded which could reflect participants' uncertainty about which utterance a speaker would use and that these individual graded ratings drive the graded average ratings. If we look at individual participants' ratings, it appears to be a combination of both. Figure 4 shows the responses of three individual participants in the *might-probably* condition. These figures show that there is a range of gumball proportions for each participant for which they assigned similar ratings to two utterances, which suggests uncertainty about the speaker's utterance choice. At the same time, however, this range also differed across participants: Participant #8, who considered the experimental speaker a "*cautious*" speaker, thought that the speaker would only be likely to use PROBABLY when the objective probability of getting a target color gumball was greater than 0.75, whereas participant #15, who considered the experimental speaker a "*confident*" speaker, thought that PROBABLY was a better utterance choice than MIGHT when the objective probability of getting a target color gumball was just greater than 0.5. These observations suggest that for some event probabilities, participants have uncertainty about a speaker's choice of uncertainty expression and that participants have a

priori different expectations about how a generic speaker would use these expressions.

This uncertainty and variability seems to be particularly borne out in the *might-probably* condition. For this reason, we chose this pair of expressions to study listeners' adaptation to variable uses of uncertainty expressions.

## 3 Modeling expectations about uncertainty expression productions

In this section we report a computational model of expectations about uncertainty expression productions that is informed by the data from the experiment reported above, and which will serve as proxy for listeners' baseline generative model of a generic speaker, and which we will use as the basis for investigating adaptation processes. What are the properties that this model should have?

Experiment 1 confirmed previous findings that participants' expectations about how a generic speaker would use uncertainty expressions depend on the set of utterances that participants can choose from. We further found that ratings were graded in part because participants seemed to have uncertainty in their expectations about how a generic speaker would use uncertainty expressions. Hence, a model predicting participants' beliefs about a speaker's productions of uncertainty expressions should (a) be able to capture differences in ratings depending on the availability of alternative utterances; (b) provide graded predictions about utterance probabilities; and (c) be able to capture within-participant uncertainty about probability of use.

Computational game-theoretic models such as the Rational Speech Act framework (RSA; Goodman and Frank, 2016) are uniquely suited to fulfill these desiderata. RSA models are a probabilistic formalization of Gricean pragmatics which model comprehension as Bayesian probabilistic inference. They consist of listener and speaker agents which recursively reason about each other to derive interpretations and choose utterances. For our purposes of modeling production expectations, we focus on the speaker model, which crucially bases its predictions on a set of alternative utterances. According to an RSA model, a speaker who wants to convey some information to a listener chooses her utterance based on the utterance's utility compared to the utility of alternative utterances. The speaker's utterance utility is determined by trading off the informativity of the utterance to a literal listener on the one hand and the cost of the utterance on the other.

In defining the informativity of an utterance, we follow previous RSA models of uncertainty expressions (Herbstritt and Franke, 2019; Lassiter and Goodman, 2017) and assume that uncertainty expressions have a threshold semantics, i.e., for each uncertainty expression $e$, there exists some threshold $\theta_e \in [0, 1]$ such that an utterance $u_e$ with $e$ is semantically felicitous if the probability $\phi$ of the proposition embedded under $e$ exceeds $\theta_e$. For example, if we assume the threshold for *might*, $\theta_{might}$, is 0.1, then the statement "It might rain this afternoon" is true if the probability of rain in the afternoon exceeds 0.1. Formally, we base the computation of informativity on a probability distribution from utterances to event probabilities $\phi$, which is usually referred to as the *literal listener* $L_0$ in the RSA framework.

$$L_0 \left( \phi \mid u_e, \theta_e \right) \propto P(\phi) 1 \left[ \phi > \theta_e \right] \quad \text{(for positive embedded propositions)}$$

$$L_0 \left( \phi \mid u_e, \theta_e \right) \propto P(\phi) 1 \left[ \phi < \theta_e \right] \quad \text{(for negated embedded propositions)}$$

$P(\phi)$ is a prior distribution over event probabilities, which is independent of the utterance by the speaker.

A *pragmatic speaker* $S_1$ who wants to communicate an event probability $\phi$ then chooses her utterance $u_e$ with uncertainty expression $e$ from a set of utterances $U$ according to a soft-max choice rule (Luce, 1959; Sutton & Barto, 1998) such that she chooses $u$ with a probability proportional to her speaker utility.

$$S_1\left(u_e \mid \phi, \theta, c\right) \propto exp\left(\lambda\left(\log L_0\left(\phi \mid u_e, \theta_e\right) - c(u_e)\right)\right)$$

$\lambda$ is a rationality parameter which governs how likely a speaker is to choose the utterance that maximizes her utility; as $\lambda$ approaches infinity, a speaker is more likely to always choose the optimal utterance.
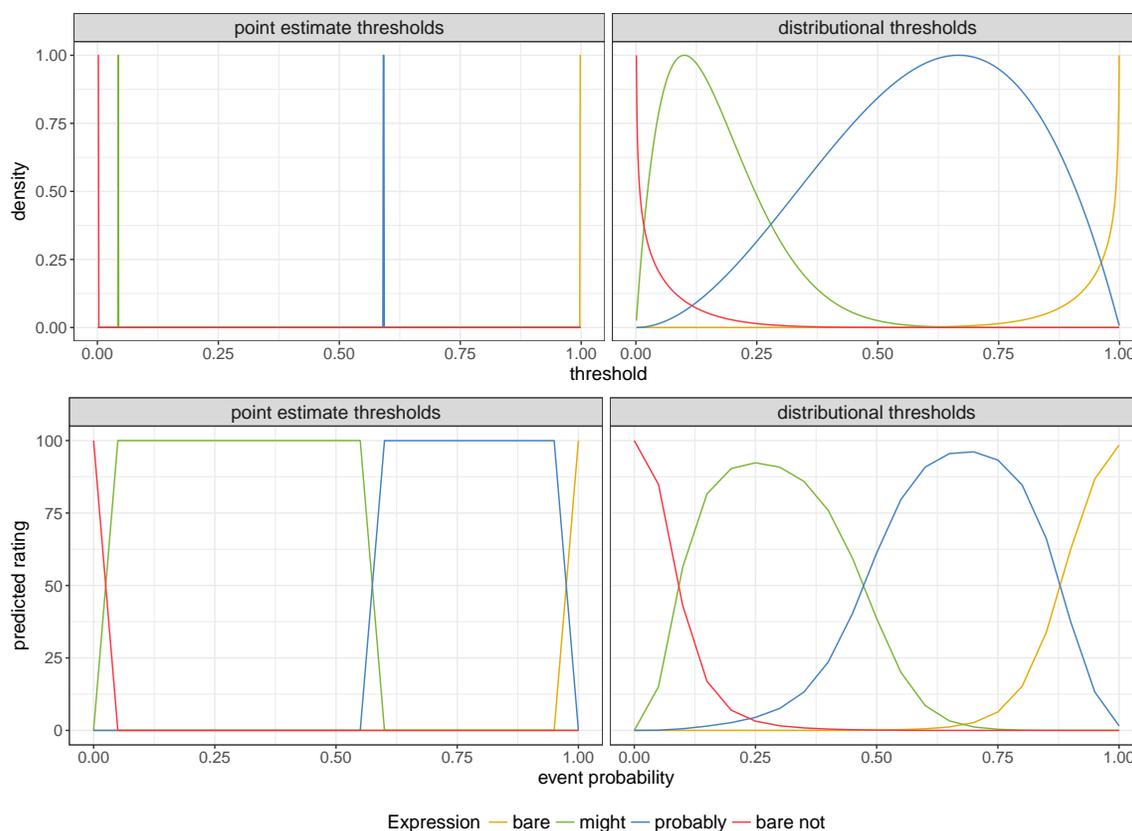


*Figure 5*. Example threshold distributions (upper panels) and corresponding model predictions by the *expected pragmatic speaker* model (lower panels). In this example, the set of possible utterances is $U = \{\text{BARE, MIGHT, PROBABLY, BARE NOT}\}$, all utterances have equal costs, the rationality parameter $\lambda$ is set to 10, and the prior probability over event probabilities $P(\phi)$ is a uniform distribution. As the panels on the left show, point estimates of thresholds lead to sharp categorical boundaries in the model predictions, whereas distributions over thresholds, as in the panels on the right, lead to gradually increasing and decreasing predicted utterance ratings.

$S_1$ crucially depends on a vector of thresholds $\theta$ which contains a threshold for each

uncertainty expression in the utterances in $U$, as well as a cost function $c(u)$. The values that speakers assign to these variables are unknown a priori; we infer these values from the data collected in the previously above reported experiment. In Experiment 1, we found that both at the population-level and at the individual-level, participants' ratings of the different expressions gradually increased and decreased with changing event probabilities (as, for example, shown in Figure 4). This is expected if we assume that participants have probabilistic beliefs about thresholds $\theta$ (as illustrated in the right panels of Figure 5) but not so if we assume that participants are reasoning based on point estimates of $\theta$ (as illustrated in the left panels of Figure 5). Considering these observations, we assume that listeners hold beliefs about speakers' thresholds in the form of a distribution $P(\theta_e)$.[7] Analogously, we assume that listeners also have beliefs $P(c)$ about the speaker's cost function. Using these two distributions, we can define the *expected pragmatic speaker* $ES_1(u_e \mid \phi)$ as follows:

$$ES_1(u_e \mid \phi) = \int P(c) \int_0^1 P(\theta) S_1(u_e \mid \phi, \theta, c)\, d\theta\, dc$$

This model predicts which utterance a listener who has uncertainty about a speaker's thresholds and cost function would expect that speaker to use to describe different event probabilities. Intuitively, this model is a weighted average of different speaker models with differing thresholds and cost functions where the weights are determined by the listener's belief distributions over thresholds and costs.

## 3.1 Linking function

We assume that in Experiment 1, participants, when asked to provide ratings for utterances, reasoned about the speaker's likely descriptions of varying event probabilities. We assume that this reasoning was guided by participants' beliefs about the speaker's thresholds and costs, and that participants averaged over their uncertainty. For this reason, we assume that the population-level average ratings of what participants expect the speaker to say in different situations correspond to the probabilities predicted by the *expected pragmatic speaker* model. Further, given the forced choice nature of the experiment and that we are estimating model parameters from limited and potentially noisy data, we make the following additional linking assumptions for which we provide a rationale and an assessment of their importance in turn.

- **Set of utterances**: Across all conditions, we assume that the set of utterances that participants are considering is the set of all utterances that we used in Experiment 1, i.e., $U = \{$ BARE, MIGHT, PROBABLY, THINK, LOOKS LIKE, COULD, BARE NOT$\}$. We include all utterances instead of only the utterances that are presented in a given condition since we assume that participants' general knowledge of English uncertainty expressions also influences their ratings. Ideally, we would include even more utterances

---

[7] We leave it open whether a speaker samples from a distribution over thresholds when making utterances (as suggested by Qing and Franke (2015)) or always uses the same values for thresholds. In the former case, listeners could have higher-order beliefs $P(\eta)$ about different speakers' threshold distributions instead of having direct beliefs about the thresholds that different speakers use. For our purposes, this distinction does not matter since we would assume that listeners marginalize over their higher-order beliefs $P(\eta)$ such that $P(\theta_e) = \int P(\eta) P(\theta_e \mid \eta)\, d\eta$ and we therefore take the simplest approach and directly model $P(\theta_e)$.

in this set of alternatives but since we can only estimate parameters for uncertainty expressions for which we collected ratings, we are limited to the utterances in $U$.

The exact set of utterances appears to be not that important for fitting the data as long as the set includes the BARE and BARE NOT utterances as well as at least one utterance with a weaker (e.g., *might*) and one with a stronger (e.g., *probably*) uncertainty expression.

- ***something else* option**: Participants in condition $\mathscr{C} = \{u_a, u_b\}$ could only choose between the three utterances $U' = \{u_a, u_b, \textit{something else}\}$. For modeling data from condition $\mathscr{C}$, we therefore need a function to predict the ratings for the utterances in $U'$. For $u_a$ and $u_b$, this is straightforward: We assume the probability of a participant choosing $u_a$ or $u_b$ is proportional to $ES_1(u_a \mid \phi)$ and $ES_1(u_b \mid \phi)$, respectively. We model the probability of a participant choosing the *something else* option as the sum of the probability of all utterances that were not part of the condition as well as a constant $O$, which accounts for probability mass assigned to utterances that participants might be considering but which are not contained in $U$. This gives us the following condition-specific function $ES_1^{(\mathscr{C})}$ for predicting participants' ratings.

$$ES_1^{(\mathscr{C})}(u \mid \phi) \propto \begin{cases} ES_1(u \mid \phi) & \text{if } u \in \mathscr{C} \\ O + \sum_{u \notin \mathscr{C}} ES_1(u \mid \phi) & \text{if } u \text{ is } \textit{something else} \end{cases}$$

This summation over alternative utterances is crucial for fitting the data since we need to capture the ratings for *something else*. The only viable alternative would be to fit individual curves for *something else* for each condition, which would require the estimation of considerably more parameters and would not explain the ratings for the *something else* option. The inclusion of the constant $O$ is less important but it still improves model fit.

- **Cost function**: We assume that the cost function represents participants' beliefs about the speaker's preferences for different utterances. Lower costs of an utterance indicate higher speaker preferences. We further assume that we are cueing participants to believe that the speaker would be likely to use the two utterances, $u_a$ and $u_b$, that are provided in condition $\mathscr{C} = \{u_a, u_b\}$ and that participants therefore primarily use the *something else* option when both of the two utterances are semantically infelicitous or otherwise highly unexpected. We model this cueing effect in our choice of the cost function $c(u)$, which depends on the condition. For the two utterances that are presented to the participants, we set the cost to 1 and for all the other utterances, we set the cost to a constant $\gamma$:

$$c(u, \mathscr{C}) = \begin{cases} 1 & \text{if } u \in \mathscr{C} \\ \gamma & \text{otherwise} \end{cases}$$

Theoretically, we could have also used a different constant $\gamma_u$ for each utterance. The data from Experiment 1, however, suggests that participants generally did not prefer one utterance over the other. To limit the number of free model parameters and to

prevent overfitting, we therefore use a single constant $\gamma$ for all utterances. We will, however, relax this assumption in our adaptation model in Section 5, which we use to investigate whether listeners update their beliefs about preferences during adaptation.

This condition-specific cost function is important for the model fit. If we didn't use such a cost function, the model would assign much higher ratings to the *something else* option than participants did.

- **Noise**: Finally, to account for participants not paying attention or making mistakes, we also include a noise term that models participants providing random ratings. The amount of noise is captured by the noise strength parameter $\delta$. This parameter indicates the proportion of random responses, that is, the proportion of responses drawn from a uniform distribution over the three condition-specific responses $U'$.

  The inclusion of the noise term is not crucial for fitting the data but it does improve model fit and is common practice in RSA models whose parameters are estimated from experimental data (see Herbstritt & Franke, 2019; Tessler & Goodman, 2019).

Incorporating all of these assumptions, we end up with the following noisy, condition-specific expected pragmatic speaker model $ES_1^{(\mathscr{C})'}(u \mid \phi)$, which we use to predict participants' ratings:

$$ES_1^{(\mathscr{C})'}(u \mid \phi) = \delta \times \frac{1}{|U'|} + (1 - \delta) \times ES_1^{(\mathscr{C})}(u \mid \phi)$$

For the prior distribution over event probabilities $P(\phi)$, which is used in the literal listener $L_0$, we use a uniform distribution over the interval $[0, 1]$.[8] For the distributions over thresholds $P(\theta_e)$, we use a Beta distribution parametrized by $\alpha_e$ and $\beta_e$. The choice of Beta distributions is motivated by two of its properties. First, the support of a Beta distribution is the interval $[0, 1]$ which corresponds to the exact range of possible values for $\theta_e$.

The second reason for using Beta distributions is that, depending on the parameterization, Beta distributions can take on very different shapes. This property is important because we are making the simplifying assumption that all utterances in our experiments have a threshold semantics. Such a semantics is commonly assumed for uncertainty expressions such as *probably* (e.g., Lassiter, 2017; Yalcin, 2010), but it is unconventional for bare assertions such as *"You'll get a blue one"*, which are generally assumed to be semantically felicitous only if the event is certain to happen, i.e., it has an event probability of 1. However, since Beta distributions can have a shape like the distribution for BARE in the upper right panel in Figure 5, the model has the capability to infer a semantics for the bare form that is almost equivalent to a traditional semantics of bare assertions. In this parameterization of the Beta distribution, most probability mass is assigned to values of $\theta$ close to

---

[8]To verify the assumption that the prior on event probabilities is uniform, we conducted a separate norming study in which participants rated how likely they thought it was that a speaker described different gumball machines containing different proportions of blue and orange gumballs after hearing an unintelligible utterance. We found that on average participants rated all gumball machines equally likely which suggests that the prior over event probabilities is indeed uniform.

1, which is mathematically almost equivalent to a traditional semantics.[9] Therefore, using Beta distributions for the threshold distributions has the desirable effect of allowing us a unified treatment of all expressions included in the model.

## 3.2 Parameter estimation

Given all the assumptions outlined above, the model has 18 parameters in total: A cost parameter $\gamma$, a rationality parameter $\lambda$, a noise strength parameter $\delta$, a constant corresponding to other utterances $O$, and for each utterance, Beta distribution parameters $\alpha_e$ and $\beta_e$. We estimated these parameters jointly from all 21 conditions of Experiment 1 using Bayesian data analysis (BDA; see, e.g., Kruschke, 2015). To construct the dataset, we treated the ratings by each participant as a probability distribution from which we sampled 10 utterances. We used highly uninformative uniform priors over the interval $[0, 15]$ for the Beta distribution and cost parameters, uniform priors over the interval $[0, 7]$ for the rationality parameter, and uniform priors over the interval $[0, 0.5]$ for $O$ and the noise strength parameter. We estimated the vector of parameters $\Theta$ using MCMC with a Metropolis Hastings sampler. To decrease autocorrelation of the chain, we collected a sample only at every 10th iteration (i.e., we use thinning of 10). We discarded the first 10,000 burn-in samples and then collected 50,000 samples. We ran four MCMC chains and confirmed convergence by computing the $\hat{R}$-statistic (Gelman, Carlin, Stern, & Rubin, 2003). More details on the implementation of the model can be found in Appendix C.
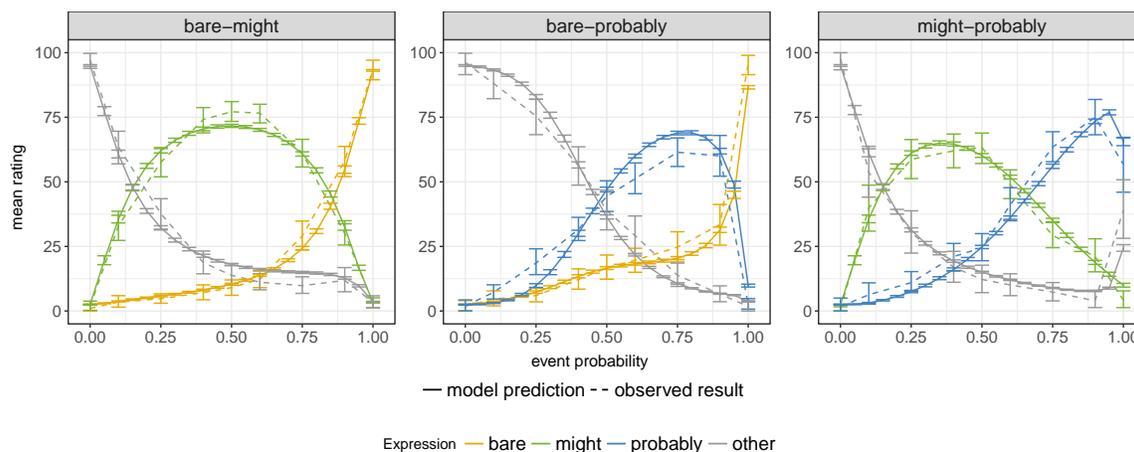
## 3.3 Model evaluation



*Figure 6*. Model predictions and results from Experiment 1. Error bars correspond to 95% high density intervals (model predictions) and bootstrapped 95%-confidence intervals (observed results).

---

[9]Alternatively, one can also see the threshold distribution for the bare form as a distribution over a verification parameter $\eta$ that governs how certain a speaker has to be to utter a bare assertion (see, e.g., Moss, 2018). Mathematically, our assumption of bare forms having a threshold semantics is equivalent to assuming that bare assertions are only semantically felicitous when a speaker's credence of the proposition exceeds the verification threshold $\eta$.

The result of the parameter estimation procedure is a posterior distribution over parameters given the observed data $P(\Theta \mid D_{obs})$. We evaluated the model fit by performing a posterior predictive check (PPC; Kruschke, 2015). To this end, we took 10,000 samples of parameters $\Theta$ from the posterior distribution and for each sample, we computed the model predictions $ES_1^{(\mathscr{C})'}(u \mid \phi)$ parameterized by $\Theta$. We then compared the average model predictions to the mean ratings that participants had provided in the pre-exposure experiments. We further computed the 95% high density interval (HDI; Kruschke, 2015) which reflects the certainty of the model about its predictions.

Figure 6 shows the model predictions and the experimental data for three conditions (see Table 1 and Appendix D for modeling results for all 21 conditions). As these plots show, the model is able to capture almost the entire variance in participants' average ratings. Further, the 95% HDIs are very small which suggests that the model is certain about its predictions. Both of these observations are also true for the model's predictions for all the other conditions. For 19 of the 21 conditions, the $R^2$ value between the model predictions and the experimental data exceeds 0.9, and for the remaining 2 conditions, the $R^2$ value exceeds 0.88.

Most cases in which the model predictions and the experimental data deviate concern the ratings at the two extremes of the event probability space. The model often underpredicts ratings for the *something else* option when there is either a 0% or a 100% chance of getting a target color gumball. In these situations, participants presumably thought that BARE and BARE NOT are the most appropriate utterances and therefore rate *something else* highly unless we provide them with the BARE or BARE NOT options. The model predicts this behavior to some extent but seems to assume that participants were cued more heavily towards the presented utterance options than they actually were. This could be an indicator that we should revisit our unconventional approach of treating the bare forms like uncertainty expressions with a threshold semantics, since the model would predict higher ratings for the *something else* at both ends of the scale if we assumed that the bare form and its negation were only true in the cases of 100% and 0% event probabilities, respectively. However, for our purposes in this paper, the exact predictions about production choices for objectively certain events are not as important and hence we decided against revising the assumption that all utterances in the model have a threshold semantics.

One potential concern given the flexibility of the model is that it could be overfitting the data. This is unlikely considering that we are estimating only 18 parameters to predict in total 567 data points (27 data points for each one of the 21 conditions) but to nevertheless rule out this possibility, we performed a leave-one-out cross-validation of the model. For each condition $x$, we estimated a distribution over parameters $\Theta_x$ using the data from all conditions but $x$. We then compared the model predictions of the model parametrized by $\Theta_x$ to participants' ratings in condition $x$. This way, the model has to predict participant behavior which it has not observed during parameter estimation. Table 1 shows the $R^2$ values for participants's ratings and model predictions for the model estimated from all conditions and the leave-one-out models.

As this table shows, the $R^2$ values remain high even if we exclude the data on which the model is evaluated from the model's training data, which suggests that our proposed model indeed explains participants' expectations of a generic speaker's uncertainty expressions.

Lastly, one of the advantages of Bayesian cognitive models is that their parameters are

| Condition | $R^2$ (all data) | $R^2$ (leave-one-out) |
|---|---|---|
| bare-might | 0.992 | 0.988 |
| bare-probably | 0.978 | 0.976 |
| bare-could | 0.978 | 0.976 |
| bare-looks like | 0.927 | 0.896 |
| bare-think | 0.968 | 0.964 |
| might-probably | 0.964 | 0.954 |
| might-could | 0.921 | 0.910 |
| might-looks like | 0.934 | 0.918 |
| might-think | 0.946 | 0.934 |
| probably-could | 0.961 | 0.959 |
| probably-looks like | 0.944 | 0.931 |
| probably-think | 0.888 | 0.860 |
| could-looks like | 0.924 | 0.910 |
| could-think | 0.931 | 0.920 |
| looks like-think | 0.970 | 0.960 |
| bare not-bare | 0.894 | 0.848 |
| bare not-might | 0.968 | 0.958 |
| bare not-probably | 0.910 | 0.893 |
| bare not-could | 0.910 | 0.840 |
| bare not-looks like | 0.927 | 0.903 |
| bare not-think | 0.933 | 0.920 |

Table 1

*$R^2$ values for experimental data and model predictions for model estimated from all data and for models estimated from all conditions except the predicted condition.*
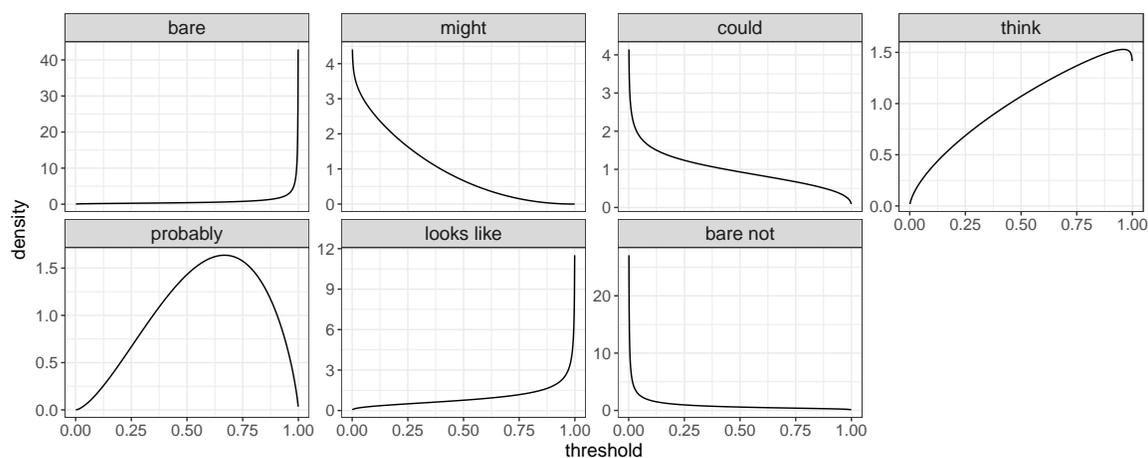


*Figure 7.* Inferred threshold distributions. For the negative bare utterance (BARE NOT), the distribution is over an upper threshold, i.e., a bare statement embedded under negation is true if the probability of the event is lower than the threshold. For all other utterances, the distribution is over a lower threshold.

interpretable. Figure 7 shows the maximum likelihood estimates of the inferred threshold distributions $P(\theta)$ for the seven uncertainty expressions that we included in our experiments. The first observation is that most threshold distributions have considerable variance rather than being peaked at a particular value. This suggests that listeners have probabilistic beliefs about the semantic thresholds.

As we discussed above, for the bare form and its negation, we expected the model to infer threshold distributions whose probability mass is concentrated around $\theta = 1$ and $\theta = 0$, respectively.[10] As Figure 7 shows, this is indeed what the inferred threshold distributions look like.

The threshold distribution for *might* has most of its probability mass concentrated at values slightly above 0. This is in line with non-probabilistic accounts of the interpretation of epistemic modals. These accounts generally assume that *might p* is true if there exists *some* world $w$ in a set of (contextually restricted) epistemically accessible worlds $E$ such that $p$ is true in $w$ (e.g., Hacquard, 2011; Kratzer, 1991; Swanson, 2008). One way to translate this logical condition into our probabilisitic framework is to assume that in our gumball machine context, there exists an epistemically accessible world $w$ for each gumball $g$ and that in world $w$, one gets gumball $g$. Under this assumption, *"You might get a blue gumball"* is true if there exists an epistemically accessible world $w$ in which one gets a gumball $g$ that is blue. At the same time, if such a world exists, then $P(\text{blue gumball})$ is greater than 0, which approximately corresponds to the threshold semantics with the inferred threshold distribution of the model. The inferred threshold distribution for *could* is similar to the one of *might*, which is again in line with non-probabilistic accounts, which assume that epistemic *might* and epistemic *could* are semantically equivalent (Hacquard, 2011; Kratzer, 1991).

The threshold distribution for *probably* has most of its probability mass concentrated at thresholds above .5. This is again compatible with existing accounts that assume that *probably p* is true if $p$ is more likely than the negation of $p$ (e.g., Kratzer, 1991). However, it is also noteworthy, that the inferred distribution has some probability mass below .5, which empirically corroborates theoretical arguments by Yalcin (2010) that *probably p* can sometimes also be true if $p$ is less likely than the negation of $p$.

The threshold distributions for the remaining expressions, *looks like* and *think* also match intuitions. The distribution for *looks like* has most of its probability mass near threshold values of 1 but is overall slightly weaker, i.e., assigns higher probabilities to lower thresholds, than the bare form. The distribution for *think* assigns most probability mass to high thresholds, which is compatible with the intuition that speakers use *think* when they strongly believe the embedded proposition but are not entirely certain that it is true.

Table 2 shows the MAP values and credible intervals for the remaining parameters. The model inferred that speakers are relatively likely to choose an optimal utterance (reflected in the $\lambda$ parameter being greater than 1); that utterances that are not included in the experiment incur a considerable cost (reflected in the $\gamma$ parameter being greater than 1); that about 7.4% of the data should be treated as noise (reflected in the $\delta$ parameter); and that the production probability of utterances not included in our set of utterances is low.

––––––––––

[10]Note that since the negation of the bare form is a negative form, $\theta$ is an upper threshold. For the bare form as well as all the other utterances that we are are considering in this paper, $\theta$ is a lower threshold.

| | $\lambda$ (rationality) | $\gamma$ (cost) | $\delta$ (noise) | $O$ (other utterances) |
|---|---|---|---|---|
| MAP | 2.21 | 3.03 | 0.074 | $3.64 \times 10^{-5}$ |
| CI | [2.16, 2.26] | [2.98, 3.08] | [0.069, 0.079] | $[2.89 \times 10^{-5}, 4.50 \times 10^{-5}]$ |

Table 2

*Maximum a posteriori estimates (MAP) and 95% credible intervals (CI) for model parameters.*

### 3.4 Interim summary

In this section, we described a computational model of production expectations of uncertainty expressions. This model is couched within the RSA framework and assumes that listeners hold beliefs about a speaker's lexicon (in the form of utterance-specific threshold distributions) and about speaker preferences (in the form of utterance-specific costs). We estimated the free parameters of this model from the results of Experiment 1, which resulted in a model that is able to accurately predict participant's utterance ratings – i.e., their expectations of use – in Experiment 1 across all conditions with a shared set of parameters.

In the following sections of this paper, we will use this model as the basis for modeling adaptation. Since this model is able to capture different beliefs about thresholds and preferences, it provides us with the opportunity to simulate the adaptation process as a result of updating beliefs about these model parameters. Further, in order to answer our primary research question of whether listeners update their beliefs about lexica or preferences, we compare different adaptation simulations in which we allow different types of parameters to be updated.

### 4 Experiment 2: Adaptation of speaker expectations

We now turn to our main research questions of whether and how listeners adapt to variable uses of uncertainty expressions. In Experiment 1, we found that participants show uncertainty in their expectations about a generic speaker's use of *might* and *probably*. Based on these results, we investigate in two experiments whether participants form speaker-specific expectations about the use of *might* and *probably*.

### 4.1 Experiment 2a

In this experiment, we test whether participants update their production expectations after observing a specific speaker's use of uncertainty expressions for a short period of time. The procedure, materials and analyses were pre-registered at https://osf.io/w926x/.

**4.1.1 Participants.** We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2 which amounted to an hourly wage of approximately $12–$15. None of the participants had previously participated in Experiment 1.

**4.1.2 Materials and procedure.**

*Exposure trials:* In the first part of the experiment, participants saw 20 exposure trials. These trials had a similar setup as the trials in Experiment 1: they also showed a child requesting a blue or orange gumball and a gumball machine with blue and orange

| | Experiment 2a | | | | | | Experiment 2b | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIGHT | | PROBABLY | | BARE | | MIGHT | | PROBABLY | | BARE | |
| | $n$ | $\phi$ | $n$ | $\phi$ | $n$ | $\phi$ | $n$ | $\phi$ | $n$ | $\phi$ | $n$ | $\phi$ |
| *cautious* | **10** | **60%** | 5 | 90% | 5 | 100% | **10** | **60%** | 10 | 90% | 5 | 100% |
| *confident* | 5 | 25% | **10** | **60%** | 5 | 100% | 10 | 25% | **10** | **60%** | 5 | 100% |

Table 3

*Number of exposure trials (n) per utterance (*MIGHT*, *PROBABLY*, *BARE*) and associated proportion of target color gumballs (ϕ) in the* cautious *vs.* confident *speaker conditions in Experiments 2a and 2b. Critical trials bolded.*

gumballs. However, instead of the cartoon adult, they showed a video of an adult male or female speaker (counterbalanced across participants) producing one of the following six utterances:

- You'll get a blue/orange one (BARE)

- You might get a blue/orange one (MIGHT)

- You'll probably get a blue/orange one (PROBABLY)

The number of trials with each of these utterances as well as the gumball proportions varied across two conditions (see Table 3 for an overview). In the *confident speaker* condition, participants saw 10 critical trials with 60% target color gumballs and the speaker producing an utterance with *probably* (target color was randomized across trials), 5 filler trials with 100% target color gumballs and the speaker producing BARE, and 5 filler trials with 25% target color gumballs and the speaker producing MIGHT. In the *cautious speaker* condition, participants saw 10 critical trials with 60% target color gumballs and the speaker producing an utterance with *might*, 5 filler trials with 100% target color gumballs and the speaker producing BARE, and 5 filler trials with 90% target color gumballs and the speaker producing PROBABLY. The filler trials contained utterance-event probability pairs that were rated very highly in the *might-probably* condition of Experiment 1 (see Figure 3) and were intended to boost confidence in the speaker.

Participants were instructed to watch what the speaker had to say to the child. The video started automatically after a 400ms delay and participants had the option to replay the video as often as they wanted. To advance to the next scene, participants had to press a button which was disabled until the video clip had finished.

***Test trials:*** The test phase was almost identical to the *might-probably* condition of Experiment 1 except that the cartoon figure of the man was replaced with a picture of the speaker that participants saw on the exposure trials. Participants were presented with scenes containing gumball machines with 9 different proportions of blue and orange gumballs (identical as in Experiment 1) and they were asked to provide ratings for the utterances MIGHT and PROBABLY by distributing 100 points across these two utterances and the blanket *something else* option. Participants provided two ratings for each of the 18 color-gumball machine combinations resulting in a total of 36 trials.

Both speakers were from the East Coast and Native Speakers of North American English. They were instructed to produce the utterances in a normal voice without any special prosody. The speakers were naïve to the purpose of the experiment.

***Attention checks.*** In order to verify that participants were paying attention to the video and the scenes, we included 15 attention checks (6 during exposure and 9 during test trials), which were randomly positioned within the two experimental phases. Trials that contained an attention check either displayed or did not display (pseudo-randomized) a small grey X somewhere around the gumball machine. After completing a trial with an attention check, participants were asked whether they had seen a grey X in the previous scenes or not.

**4.1.3 Exclusions.** We excluded participants who provided incorrect responses to more than 3 of the attention checks. Based on this criterion, we excluded 11 participants in the *confident speaker* condition and 8 participants in the *cautious speaker* condition. None of the results reported below depend on these exclusions.

**4.1.4 Analysis and predictions.** Intuitively, we expect a more confident speaker to use lower thresholds for *probably* and *might* than a more cautious speaker. Therefore, if participants track these different uses, we expect their ratings to depend on how the speaker used uncertainty expressions during the exposure phase. Concretely, in our forced choice production paradigm, we expect participants in the *confident speaker* condition to rate PROBABLY highly for a larger range of event probabilities than participants in the *cautious speaker* condition. Following Yildirim et al., 2016, we quantified this prediction by fitting a spline with four knots for each expression and each participant and computing the area under the curve (AUC) for the splines corresponding to each expression and participant. The area under the curve is proportional to how highly and for how large of event probabilities participants rate an utterance. If an utterance is rated highly for a larger range of event probabilities, the AUC will also be larger. We therefore tested whether listeners updated their expectations according to these intuitions by computing the difference between the AUC of the spline for MIGHT and of the spline for PROBABLY for each participant. We predicted that the mean AUC difference would be larger in the *cautious speaker* condition than in the *confident speaker* condition.

**4.1.5 Results and discussion.** Figure 8 shows the mean ratings for the three options in the two conditions. As these plots show, participants updated their expectations about the speaker's language use and therefore made different predictions about how the speaker would use uncertainty expressions. In the *cautious speaker* condition, participants gave high ratings for MIGHT for a larger range of event probabilities than in the *confident speaker* condition. On the other hand, participants gave high ratings for PROBABLY for a larger range of gumball proportions in the *confident speaker* condition than in the *cautious speaker* condition. These differences result in a significantly larger AUC difference in the *cautious speaker* condition than in the *confident speaker* condition ($t(59) = 4.98$, $p < 0.001$, see also left panel of Figure 9).

As Figure 8 shows, participants also differed in their ratings of the two utterances when they were presented with a scene with 60% target color gumballs. In the *cautious speaker* condition, participants rated MIGHT higher than PROBABLY; in the *confident speaker* condition, the pattern was reversed and participants rated PROBABLY higher than MIGHT. These expectations mirror the speaker behavior during the exposure phase and provide
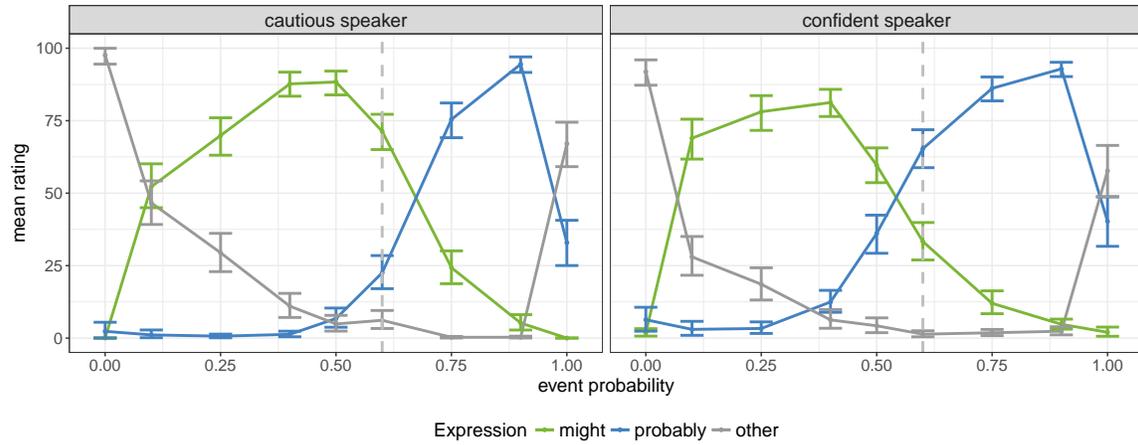
*Figure 8*. Mean post-exposure ratings from Experiment 2a. Error bars correspond to bootstrapped 95%-confidence intervals. The grey dotted line highlights the ratings for the 60% event probability ratings.
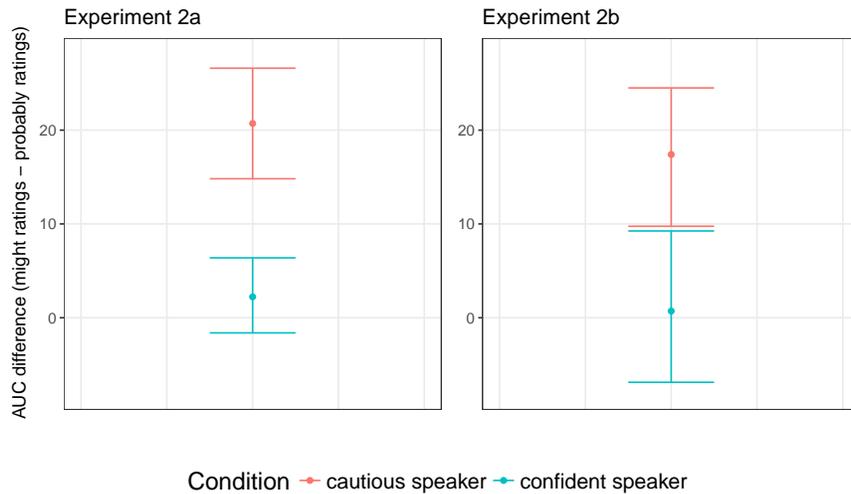


*Figure 9*. Area under the curve (AUC) differences from Experiments 2a and 2b. Error bars correspond to bootstrapped 95%-confidence intervals.

additional evidence that participants tracked the speaker's usage of uncertainty expressions.

Our results further suggest that participants updated their mappings between uncertainty expressions and event probabilities: In the *confident speaker* condition, MIGHT and PROBABLY were rated highly for lower event probabilities than in the *cautious speaker* condition. However, one potential confound in this experiment is that the number of utterances with *might* and *probably* differed across the two conditions (see left part of Table 3). It is possible that participants only learned that the *cautious speaker* overall prefers to use MIGHT and the *confident speaker* prefers PROBABLY. To address this confound, we conducted another experiment in which we balanced the number of exposures to MIGHT and PROBABLY.

### 4.2 Experiment 2b

In this experiment, we aimed to replicate the results from Experiment 2a and test whether listeners update their mappings between the uncertainty expressions and event probabilities when they are exposed to a speaker who uses *might* and *probably* with the same frequency. The procedure, materials and analyses were pre-registered at https://osf.io/qnam7.

**4.2.1  Participants.**  We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2.20 which amounted to an hourly wage of approximately $12–$15. None of the participants had participated in any of the previous experiments.

**4.2.2  Materials and procedure.**  Materials, conditions, and procedure were identical as in Experiment 2a except that we added 5 additional fillers such that the frequency of MIGHT and PROBABLY was the same (10 utterances per expression) in both conditions. See right part of Table 3 for an overview.

**4.2.3  Analysis and predictions.**  The analysis was identical as in Eperiment 2a. We again predicted that the mean AUC difference would be bigger in the *cautious speaker* condition than in the *confident speaker* condition.

**4.2.4  Exclusions.**  We used the attention checks and exclusion criteria as in Experiment 2a. Based on these criteria, we excluded 8 participants in the *cautious speaker* condition, and 7 participants in the *confident speaker* condition.

**4.2.5  Results and discussion.**  Figure 10 shows the mean ratings for the three options in the two conditions. As these plots show, participants updated their expectations about the speaker's language use. As in Exp. 2a, the AUC difference was bigger in the *cautious speaker* condition than in the *confident speaker* condition ($t(63) = 2.99$, $p < 0.01$, see also right panel of Figure 9).

This experiment provides evidence against an account according to which participants only learn that the *cautious speaker* prefers to use MIGHT and the *confident speaker* prefers PROBABLY. Since the frequency of both expressions was the same in this experiment, participants could not have inferred a preference for one of these two utterances. That we nevertheless observed different ratings in the two speaker conditions suggests that participants updated their beliefs about the mapping between expression and event probabilities.

In summary, the results from Exps. 2a and 2b provide evidence for listener adaptation to a specific speaker's use of uncertainty expressions after a brief exposure phase. Further,
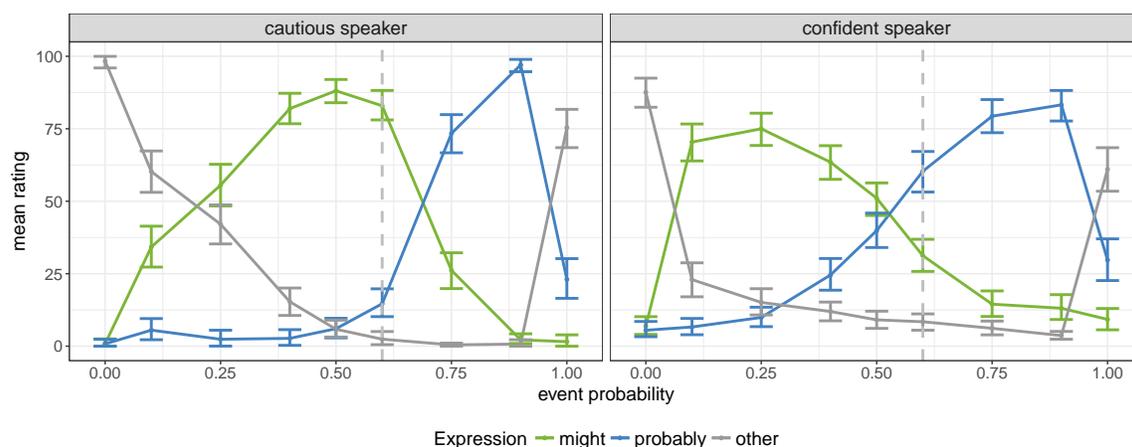
*Figure 10*. Mean post-exposure ratings from Experiment 2b. Error bars correspond to bootstrapped 95%-confidence intervals. The grey dotted line highlights the ratings for the 60% event probability ratings.

taken together, these experiments suggest that listeners' expectations about a speaker's language use are at least not exclusively driven by tracking speakers' preferences for different utterances. We investigate the nature of the updated expectations in the next section.

## 5   Adaptation model

The experimental results presented in the previous section suggest that listeners update *some* expectations about language use when they interact with a speaker. However, the nature of the updated representations is unclear. As mentioned in the introduction, there are three likely candidates: first, it is possible that listeners update their expectations about the speaker's lexicon (i.e., the mapping between event probabilities and uncertainty expressions); second, listeners might update their expectations about the speaker's preferences; and third, they might update both their expectations about the speaker's lexicon and about the speaker's preferences. The experimental results above suggest that it is unlikely that listeners track only speaker preferences, but considering that beliefs about preferences and beliefs about the lexicon can interact in complex ways (as illustrated in Figure 1), we investigate all three options.

The production expectation model in Section 3 provides us with the opportunity to formally evaluate these three hypotheses. Through a series of simulations of the adaptation process, we can compare models in which different types of parameters are updated during adaptation. Following work in modeling adaptation in other linguistic domains (e.g., Hawkins et al., 2017; Kleinschmidt, Fine, & Jaeger, 2012; Kleinschmidt & Jaeger, 2015; Qing, 2014; Roettger & Franke, 2019), we assume that in interaction, listeners form beliefs about a set of speaker-specific parameters $\Theta_S$.[11] We further assume that the formation of

---

[11]Since the manipulation in our experiments was between subjects, our results do not provide direct evidence that listeners are indeed adapting to speakers (as compared, for example, to the general experimental situation). For now, we assume that listeners are adapting to specific speakers and we return to this issue in the general discussion.

these beliefs is an instance of Bayesian belief updating: listeners start off with prior beliefs about $\Theta_S$ based on their general knowledge about language and subsequently update their beliefs about $\Theta_S$ with every utterance they hear. That is, after observing a series of productions $D = d_1, ..., d_n$ where each $d_i$ is an utterance-event probability pair $d_i = (u_i, \phi_i)$, listeners' beliefs about $\Theta_S$ are the result of performing Bayesian inference:

$$P(\Theta_S \mid D) \propto P(\Theta_S)P(D \mid \Theta_S) = P(\Theta_S) \prod_{i=1}^{n} P(d_i \mid \Theta_S)$$

We assume that the likelihood function is the *expected pragmatic speaker $ES_1$* parameterized by $\Theta_S$:

$$P(\Theta_S \mid D) \propto P(\Theta_S) \prod_{i=1}^{n} ES_1(u_i \mid \phi_i, \Theta_S)$$

## 5.1 Simulations

In order to investigate which parameters are updated during adaptation, we ran simulations with varying prior structures, which correspond to different assumptions about which parameters may be updated. The adaptation model crucially relies on a prior over speaker-specific parameters $P(\Theta_S)$ which reflects listeners' prior beliefs about the use of uncertainty expressions. For our simulations, we assumed that the means of this prior are given by the estimates of the model parameters that we obtained from fitting the model to the norming data. The variances reflect whether or not the parameter can be updated in response to exposure. In particular, we used delta distributions, i.e., a distribution with zero variance, to model a parameter that cannot be updated. We ran simulations on models with the following three prior structures:

- **Costs**: The first prior structure corresponds to an adaptation process according to which participants only learn speaker preferences during adaptation. We modeled the prior over cost parameters as a log-normal distribution centered at the mean value inferred from the norming data. Because we were interested in whether listeners update their beliefs about speaker preferences, we relaxed the assumption from the norming data model that all utterances have the same cost and assumed that each expression has its own cost parameter indicating beliefs about the speaker's preferences. Use of the log-normal distribution was motivated by two reasons: First, cost must be greater than zero, and the support of log-normal distributions is limited to numbers greater than 0. Second, since the cost term is part of an exponent in the expected pragmatic speaker model, differences on a logarithmic scale correspond to linear differences in the model's utterance probabilities. For the priors over all other parameters, we used a delta distribution.

- **Threshold distributions**: This prior structure corresponds to an adaptation process according to which participants only learn speaker-specific threshold distributions during adaptation. We parameterized threshold Beta distributions $P(\theta_e)$ with their mean $\mu_e$ and population parameter $\nu_e$ (Kruschke, 2015). Since the threshold and therefore also the mean of the threshold distribution have to lie within the interval

|                      | Range         | Step size | MAP value |
|----------------------|---------------|-----------|-----------|
| Variance for $\mu$   | [0.025,0.2]   | 0.025     | 0.175     |
| Scale for $\nu$      | [0.5,3.5]     | 0.5       | 2.5       |
| Variance for cost    | [0.1,1.5]     | 0.2       | 1.3       |

Table 4

*Explored hyperparameter ranges for variance parameters, and inferred MAP values, which were used in the adaptation simulations.*

[0,1], we used a truncated normal distribution $\mathscr{N}_{[0,1]}$, which we centered at the mean value from the norming data. For the population parameters $\nu_e$, which indicate how peaked a threshold distribution is, we assumed that distributions can only become more peaked when listeners are exposed to a speaker with very consistent language use and therefore modeled the prior as an exponential distribution shifted to the mean population parameter that we estimated from the norming data. We used a delta distribution for the priors over all other parameters.

- ***Threshold distributions and costs***: This prior structure corresponds to an adaptation process according to which participants learn both speaker-specific threshold distributions and speaker preferences during adaptation. We used the log-normal distributions as priors over the cost parameters and the truncated normal and exponential distributions as priors over the threshold distribution parameters, as described above. This means that both the threshold distributions and the cost parameters could be updated during the adaptation simulations.

Each of these prior structures corresponds to a different hypothesis about which expectations listeners update during adaptation. For comparison, we also considered a baseline in which none of the parameters are updated during adaptation (the *fixed* prior structure). To adjudicate between these three hypotheses, we ran simulations of the adaptation process for both (*cautious speaker* and *confident speaker*) conditions with different prior structures and compared the models in terms of their likelihood of generating the experimental data. During each simulation, we performed Bayesian inference to infer the posterior parameter distribution after observing the 20 data points that participants observed in the exposure phase (see Table 3 for an overview of the 20 utterances in the two conditions). We performed inference using MCMC with a Metropolis-Hastings sampler. We used thinning of 10, discarded the first 2,000 burn-in samples and collected 10,000 samples from each of the two chains.

The prior distributions over the different parameters that may be updated during the adaptation simulations are all parameterized by two constants which govern their mean and their variance. The first set of parameters (the mean of the log-normal and truncated normal distributions; the location parameter of the exponential distributions) was given by the estimates from fitting the model to the norming data. The second set of parameters (the variance of the log-normal and truncated normal distributions; the scale parameter of the exponential distributions) was treated as hyperparameters of the simulations. To keep the model as simple as possible, we only used three hyperparameters in total: a variance parameter for the cost for all expressions; a variance parameter for the mean of the

| Model | $R^2$ | odds |
|---|---|---|
| fixed | 0.746 | $10^{-1200}$ |
| cost | 0.770 | $10^{-448}$ |
| threshold distributions | 0.874 | $10^{-284}$ |
| cost & threshold distributions | 0.815 | 1 |

Table 5

*Model evaluation results on data from Experiment 2a. $R^2$ are computed between the mean post-exposure ratings and the mean model predictions. odds are the posterior likelihood odds of the models compared to the cost and threshold distributions model.*

threshold distributions for all utterances; and a scale parameter for the prior over population parameters for all utterances. We optimized these three parameters through a Bayesian hyperparameter search on the adaptation data, which provided us with a distribution over hyperparameter values. Considering that each simulation is computationally expensive, we could only test a few hundred hyperparameter combinations, which are listed in Table 4. We found that the resulting distributions were highly peaked and therefore, we used only the maximum a posteriori estimates of the hyperparameters (also shown in Table 4) for the model comparisons below.

## 5.2   Model comparisons

We compared model fits according two metrics. First, we considered the $R^2$ value between participants' average post-exposure ratings and the maximum a posteriori predictions of the post-exposure model. Second, we computed the likelihood of the model generating the post-exposure data. For the latter, we constructed a dataset $D_{obs}$ of utterance-event probability pairs by treating each post-exposure rating as a probability distribution and sampling 10 utterances from it. We then computed the posterior likelihood odds between Model 1 with posterior distribution over parameters $P(\Theta_S^{(1)})$ and Model 2 with posterior distribution $P(\Theta_S^{(2)})$.

$$\text{posterior likelihood odds} = \frac{\int_0^1 P\left(\Theta_S^{(1)}\right) P\left(D_{obs} \mid \Theta_S^{(1)}\right) d\Theta_S^{(1)}}{\int_0^1 P\left(\Theta_S^{(2)}\right) P\left(D_{obs} \mid \Theta_S^{(2)}\right) d\Theta_S^{(2)}}$$

The posterior likelihood odds indicate how much more likely it is that the data was generated by Model 1 than by Model 2. Since we are marginalizing over a distribution over parameter values, this comparison of models will naturally favor simpler models. For a more complex model with more parameters, the distribution over different parameter values will be more dispersed and can contain more parameter configurations that lead to a lower likelihood of the data.

Table 5 shows the $R^2$ values between the the models and the experimental data from Experiment 2a as well as the posterior likelihood odds. As the values in this table show, the model in which the cost as well as the threshold distributions are updated during adaptation is much more likely to generate the experimental data than the other two less complex models. However, this is not entirely reflected in the $R^2$ values between the mean

| Model | $R^2$ | odds |
|---|---|---|
| fixed | 0.662 | $10^{-1248}$ |
| cost | 0.802 | $10^{-458}$ |
| threshold distributions | 0.877 | $10^{-212}$ |
| cost & threshold distributions | 0.875 | 1 |

Table 6

*Model evaluation results on data from Experiment 2b. $R^2$ are computed between the mean post-exposure ratings and the mean model predictions. odds are the posterior likelihood odds of the models compared to the cost and threshold distributions model.*

model predictions and the empirical means. Similarly to the posterior odds, the $R^2$ values suggest that the *fixed* model and the *cost* model make worse predictions than the other two models. At the same time, the two metrics disagree on the ranking of the *threshold distributions* and *cost and threshold distributions* models – the $R^2$ values suggest that the model according to which only the threshold distributions are updated during adaptation predicts participants' post-exposure behavior best.

To assess the stability of these results, we conducted another series of simulations to predict the post-adaptation ratings from Experiment 2b. For these simulations, we used identical prior structures and parameterizations as in the previous simulations, i.e., we did not optimize any hyperparameters of the model. However, since participants saw additional 5 filler utterances during the exposure phase, we also exposed the model to 5 additional utterances. The model evaluation results for these simulations are shown in Table 6. As the likelihood odds in this table show, the model in which both the costs and the threshold distributions can be updated is again much more likely to generate the experimental data than the other two models. This replicates the findings from the previous simulations. According to the $R^2$ metric, the *threshold distributions* and *cost and threshold distributions* models predict the data approximately equally well.

What do these modeling results tell us about the semantic/pragmatic adaptation process? We assumed that each of these simulations correspond to an adaptation process in which different types of expectations are updated. The modeling results strongly corroborate the experimental results from Experiments 2a and 2b: the models according to which no expectations are updated during adaptation (the *fixed* model) or according to which only preferences are updated (the *cost* model) provide poor predictions for the post-adaptation ratings. The results also clearly indicate that listeners update expectations about the threshold distributions. Independent of the metric, the models according to which listeners update threshold distributions were best at predicting post-adaptation behavior in all simulations.

However, the conflict between the $R^2$ metric and the log odds ratio leaves unclear whether adaptation is a result of *only* updating expectations about threshold distributions, or whether update on preferences is also necessary. In part, this inconsistency can most likely be explained by the properties of the $R^2$ metric. For one, unlike the posterior odds, it does not take uncertainty of the model predictions into account but rather compares the mean participant behavior to the mean model predictions. Considering that the model does exhibit considerable uncertainty in its post-exposure rating predictions, it is not particularly
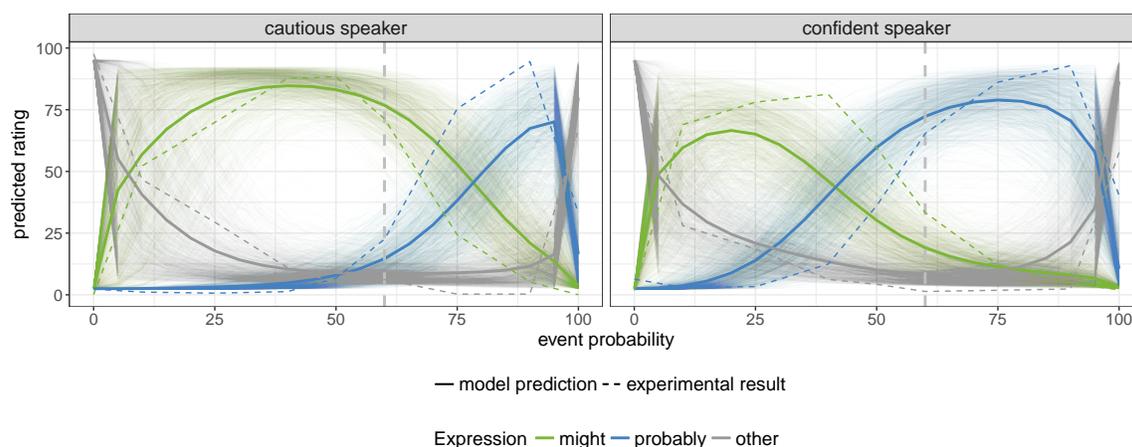
*Figure 11*. Post-adaptation model predictions from simulations for Experiment 2a and experimental results. The solid lines shows the mean model predictions and the thin lines around the mean show the distribution of model predictions.

surprising that the two metrics suggest different conclusions.[12] Second, since the ratings and the model predictions are probability distributions, the empirical and predicted ratings for one utterance for a specific event probability are not independent of the ratings for other utterances and therefore not all assumptions for using $R^2$ are met. Considering these factors, the posterior odds are to be trusted more than the $R^2$ values. This suggests that there is strong evidence that listeners update expectations about both threshold distributions and preferences.

### 5.3    Model evaluation

Apart from quantitatively assessing the fit of the model, it is informative to visually inspect the predictions of the model to verify that the model makes correct qualitative predictions. Figure 11 shows the post-exposure predictions of the *cost & threshold distributions* model compared to the average participant ratings for the two conditions from Experiment 2a.[13] Qualitatively, the model captures several important patterns in the post-adaptation behavior. The model correctly predicts that in the *cautious speaker* condition, ratings for MIGHT are higher than ratings for PROBABLY when there is an objective probability of 0.6. For the *confident speaker* condition, the model correctly predicts the opposite pattern. The model also predicts that in the *cautious speaker* condition, participants rate MIGHT highly for a larger range of event probabilities than in the *confident speaker* condition and the model predicts the reverse pattern for the PROBABLY ratings. Further, the model predicts that high ratings for MIGHT and PROBABLY are not limited to the utterance-event probability combinations that participants observed during the exposure phase. For example, the model correctly predicts high ratings of MIGHT for low event probabilities in the *cautious speaker* condition despite the fact that it never observed utterances for low

---

[12]See Gelman, Goodrich, Gabry, and Vehtari (2019) for a proposal of a Bayesian $R^2$ value.

[13]We only discuss the simulations for Experiment 2a in this section. See Appendix E for the same plots for the simulations for Experiment 2b.
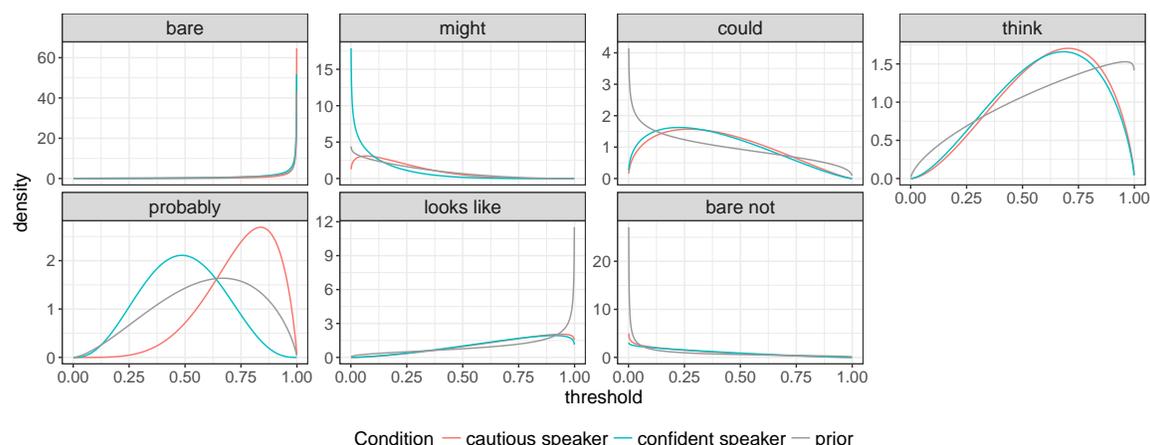
*Figure 12*. Post-adaptation threshold distributions from the simulations for Experiment 2a.

event probabilities. Similarly, the model predicts high rating of PROBABLY for high event probabilities in the *confident speaker* condition – a combination which was again not part of the exposure trials of this condition.

Quantitatively, there are some differences between the model predictions and participant behavior. This is not surprising considering that the model predictions are a result of simulations and, with the exception of the three variance parameters of the prior distributions, we did not fit any model parameters to the behavioral data. One difference is that the model underpredicts the ratings of one of the filler utterances in both conditions: in the *cautious speaker* condition, the model underpredicts ratings of PROBABLY; in the *confident speaker* condition, it underpredicts ratings of MIGHT. One reason for this deviation could be the relatively simple prior structure. For the priors, we made the assumption that all model parameters are independent of each other and that the variance for the different parameter types is the same for all utterances. However, it could be that listeners have more structured prior beliefs such that priors over different parameters are correlated or variances of prior distributions differ. For example, it could be that listeners expect the thresholds for *might* and *probably* to be correlated such that higher thresholds for *might* are correlated with higher thresholds for *probably*. Or it could be that listeners expect more between-speaker variation for some expressions than for others. Considering that we only have data from two experiments to test model predictions and therefore would likely overfit to the data if we tried to fit more complex prior structures with more parameters, we leave the investigation of the exact structure of listeners' prior beliefs to future work.

The second noticeable deviation is that the model overpredicts the ratings of the OTHER utterance for event probabilities of 1. This prediction is primarily driven by high values for the predicted ratings of BARE. However, we argue that the model predictions in this case are reasonable, and that the lower participant ratings are likely an artifact of the experimental task. After completing the experiment, several participants indicated in a feedback form that they were confused by the lack of an option to rate the BARE utterance, which they had heard during the exposure phase. In light of this confusion, almost all individual participants chose among two strategies when there was an event probability of
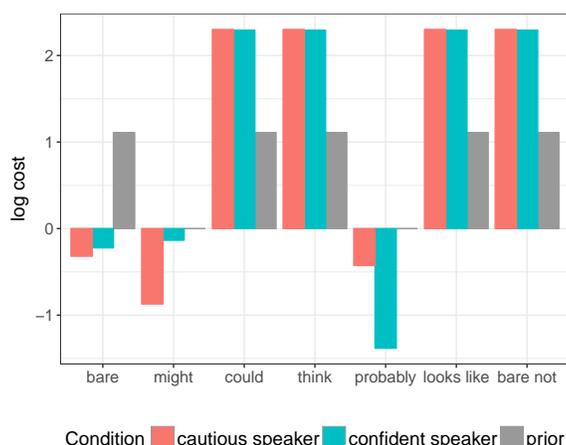
*Figure 13*. Post-adaptation *log* cost values from simulations for Experiment 2a. Note that the cost of MIGHT and PROBABLY in the norming data model was 1 and therefore the *log* cost for these utterances is 0.

1: they either provided a rating of 100 for OTHER or they provided a rating of 100 for PROBABLY, which on average leads to the ratings shown in Figure 11.

With the exception of these two deviations, the model makes not only correct qualitative, but also accurate quantitative predictions for the post-exposure ratings.

Lastly, we can also inspect how the model arrived at its predictions by looking at the inferred model parameters. Figures 12 and 13 show the inferred post-exposure threshold distributions and costs for the two conditions as well as the distributions inferred from the norming data. Figure 12 shows that the threshold distribution for *probably* changed considerably depending on the exposure phase: in the *cautious speaker* condition, its mean shifted to a higher value than inferred from the norming data; in the *confident speaker* condition the mean shifted to a lower value. To a lesser extent, we observe similar shifts in the mean of the threshold distributions for *might*. We further observe that for all expressions, the variance of the threshold distributions decreased as a result of adaptation. In the case of the expressions that were part of the exposure phase, this is expected, since the exposure speaker used these expressions very consistently; in the case of the other expressions, this decrease in variance is a result of the exponential prior over the population parameter, which biased the model towards lower variance. For some of the thresholds, this resulted in differently shaped distributions. However, note that the area under the curve of all threshold distributions except for *probably* is still very similar to the area under the curve of the norming data threshold distributions. And overall, except for the distributions for *might* and *probably*, the post-exposure threshold distributions are almost identical in both conditions. This suggests that the post-adaptation expectations are in part a result of updated threshold distributions for *might* and *probably*.

Figure 13 shows that the costs of the MIGHT, PROBABLY and BARE utterances, i.e., the three utterances that participants observed during the exposure phase, all decreased while the costs of the other four utterances increased compared to the costs inferred from the norming data. Further, the post-exposure cost of *might* is lower than the cost of *probably* in the *cautious speaker* condition and the opposite relation between these costs holds

in the *confident speaker* condition. These cost differences are expected considering that the number of exposure trials across the two conditions differed (in the *cautious speaker* condition in Experiment 2a there were more trials with MIGHT; in *confident speaker* condition more trials with PROBABLY). More surprisingly, this pattern persisted in the simulations for Experiment 2b in which the exposures of these two utterances were balanced across conditions (see Appendix E for cost plot for balanced simulations). These persistent differences suggest that the post-adaptation expectations are in part also a result of updated beliefs about preferences of MIGHT and PROBABLY.

### 5.4   Interim summary

In the previous two sections, we presented the results from two experiments, which provide strong evidence for listeners updating expectations about a speaker's use of uncertainty expressions after brief exposure to that speaker.

We further presented a computational adaptation model which models the adaptation process as an instance of Bayesian belief updating. We used different implementations of that model to investigate which kind of expectations listeners update during adaptation. We found strong evidence that listeners update beliefs about the threshold distributions and we found some evidence that listeners also update beliefs about speaker preferences.

### 6   Experiment 3: Effect of adaptation on interpretation

Up to this point, we focused on listeners' expectations about a speaker's use of uncertainty expressions. As we discussed in the introduction, we expect updated expectations to also have an effect on the *interpretation* of uncertainty expressions. This effect is also predicted by RSA models, which assume that a pragmatic listener $L_1$ tries to infer the state of the world (in our case, the event probability $\phi$) by reasoning about their prior beliefs about the world state and their expectations about a speaker's language use (in our case, the expected pragmatic speaker $ES_1$) via Bayes' rule:

$$L_1(\phi \mid u_e) \propto P(\phi)ES_1(u_e \mid \phi).$$

According to such a model of interpretation, the shifts in expectations that we observed in the previous experiment should also lead to a shift in interpretations. If we assume a uniform prior over event probabilities,[14] then the model predicts that listeners who were exposed to a *cautious* speaker should infer higher event probabilities when hearing MIGHT or PROBABLY than listeners who were exposed to a *confident* speaker. Figure 14 shows the distribution over event probabilities after hearing three different utterances as predicted by $L_1$ parameterized by the inferred parameters from our adaptation simulations in the previous section. As these plots show, in the *cautious speaker* condition, the distribution over event probabilities after hearing *might* and *probably* is shifted towards higher values as compared to the distributions in the *confident speaker* condition.

---

[14]To reiterate, this assumption was motivated by the study reported in Footnote 8, which suggested that participants on average assign equal probability to each gumball machine a priori.
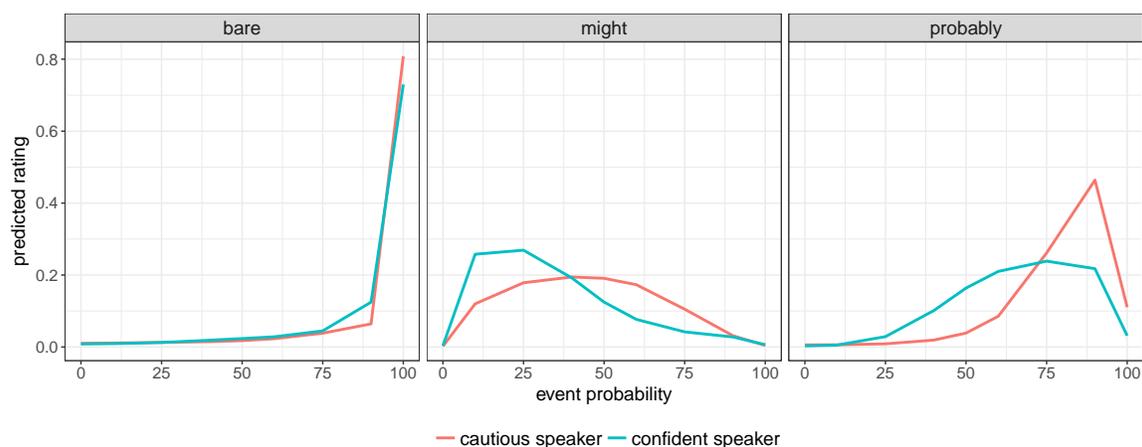
*Figure 14*. Post-adaptation interpretation distributions for the utterances BARE, MIGHT, and PROBABLY as predicted by the pragmatic listener $L_1$.

In this experiment, we tested whether this prediction is correct and whether listeners' change in expectations transfers to a change in interpretations. The procedure, materials and analyses were pre-registered at https://osf.io/ghnc3.[15]

## 6.1 Participants

We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $1.5 which amounted to an hourly wage of approximately $15. None of the participants had participated in any of the previous experiments.

## 6.2 Materials and Procedure

Participants completed a set of exposure trials followed by a set of test trials. The exposure trials were identical to the exposure trials in Experiment 2b. The test trials probed participants' interpretations of the utterances MIGHT, PROBABLY and BARE. On each test trial, participants listened to a recording of the speaker from the exposure phase producing MIGHT, PROBABLY and BARE and then participants were asked to rate for 9 gumball machines with the same proportions of blue and orange gumballs as in the previous experiments how likely they thought it was that the speaker saw each of these gumball machines by distributing coins. Participants distributed 10 coins per trial and completed 6 trials in total – one for each expression-color pair. The exposure phase again contained 6 attention check as in the previous experiment. However, given the low attention check performance in the previous experiments, we modified the attention checks. Instead of asking participants whether they saw an X on the previous trial, we asked participants to choose the gumball machine that they had seen on the previous trial among two machines displayed in random order.

---

[15]This experiment is a modified version of a previous experiment, which qualitatively yielded the same results but also seemed to confuse many participants. See Appendix F for a discussion of the original experiment.
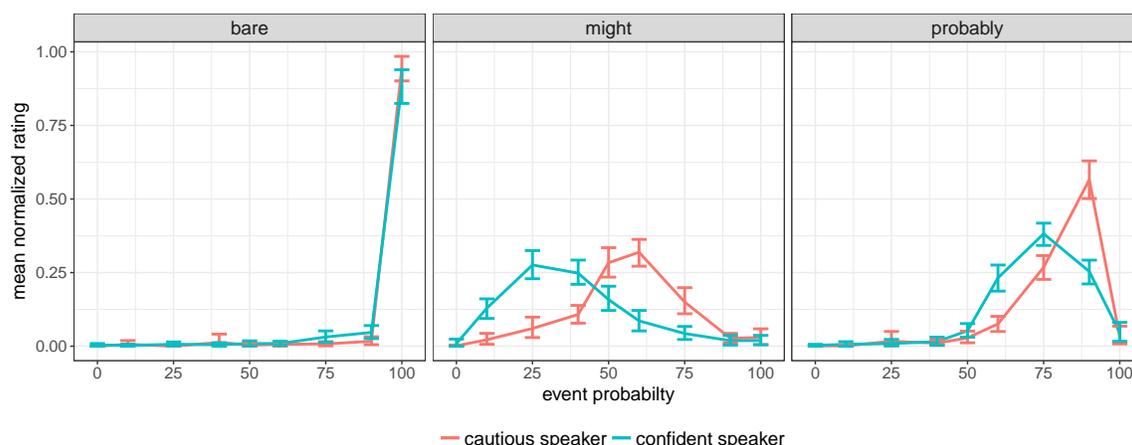
*Figure 15*. Aggregated post-exposure ratings from Experiment 3. Error bars correspond to bootstrapped 95%-confidence intervals.

### 6.3   Exclusions

We excluded participants who failed more than 2 attention checks, which led to 1 exclusion in the *cautious speaker* condition and 1 exclusion in the *confident speaker* condition.

### 6.4   Analysis and Predictions

If participants update their expectations of a specific speaker's use of uncertainty expressions, we expect that listeners interpret a more confident speaker's utterance to communicate a lower event probability than a more cautious speaker's utterance. We tested this prediction by treating participant's distributions of coins of gumball machines as a probability distribution over gumball proportions (and consequently event probabilities). For each utterance, we normalized participants' coin distributions such that they summed up to 1, so that we could interpret the normalized scores as a categorical probability distribution over gumball machines given an utterance. We computed the expected value of target color gumballs from these probability distributions and compared these expected values across the two conditions with a *t*-test. We predicted that the expected values of MIGHT and PROBABLY would be larger in the *cautious speaker* condition than in the *confident speaker* condition.

### 6.5   Results and Discussion

Figure 15 shows the aggregated and normalized ratings for the two conditions. As predicted, participants provided higher ratings for gumballs with higher target color percentages after hearing MIGHT and PROBABLY in the *cautious speaker* condition than in the *cautious speaker* condition. This also led to a significantly higher expected value for MIGHT ($t(76) = 5.84$, $p < 0.001$) and PROBABLY ($t(76) = 3.92$, $p < 0.001$) in the *cautious speaker* condition as compared to the *confident speaker* condition.

These results suggest that listeners not only update their expectations about a speaker's use of uncertainty expressions, but also use those updated expectations in interpretation.

| Model | $R^2$ | odds |
|---|---|---|
| fixed | 0.662 | $10^{-442}$ |
| cost | 0.885 | $10^{-217}$ |
| threshold distributions | 0.888 | $10^{-116}$ |
| cost & threshold distributions | 0.927 | 1 |

Table 7

*Model evaluation results on data from Experiment 3. $R^2$ are computed between the mean post-exposure ratings and the mean model predictions. odds are the posterior likelihood odds of the models compared to the cost and threshold distributions model.*

## 6.6 Model comparison

We return again to our main research question regarding which expectations are updated during adaptation. The production expectation experiments and model simulations provided strong evidence for listeners updating their beliefs about the threshold distributions. On the other hand, the two evaluation metrics provided conflicting results regarding whether or not beliefs about speaker preferences are also updated. We therefore also compared the pragmatic listener $L_1$ predictions from the simulations with different prior structures to the post-exposure ratings in Experiment 3. To this end, we computed the predictions of the $L_1$ model from the posterior distributions over model parameters that we obtained through the simulations in the previous section. Table 7 shows the model fit for the different types of simulations. As this table shows, the model according to which both threshold distributions and costs are updated provides the best fit according to both metrics. Considering that the posterior likelihood odds consistently favored this model in all three model comparisons, we take these results together as strong evidence that listeners update their expectations about threshold distributions and costs.

## 6.7 Model evaluation

Figure 16 superimposes the model predictions and the experimental data. As these plots show, the model accurately captures most of the qualitative and quantitative patterns. First, the model makes both qualitatively and quantitatively accurate predictions for the interpretation of the BARE utterance in both conditions. Second, the model makes the crucial qualitative prediction that participants expect the speaker to communicate lower event probabilities in the *confident speaker* condition than in the *cautious speaker* condition, which we also observed in Experiment 3. Further, even though we used the parameters that we obtained in the simulations from the previous section and did not fit any parameters to the data from Experiment 3, the model also provides good quantitative predictions of participant's interpretation of MIGHT and PROBABLY, which provides further support for the hypothesis that semantic/pragmatic adaptation is an instance of Bayesian belief updating.

The only main deviation between the model predictions and the experimental data lies in the interpretation of MIGHT in the cautious speaker condition. For this interpretation, the model predicts a less peaked distribution than the empirical distribution. One explanation for this deviation could be that participants are considering alternative uncertainty expressions (e.g., *very unlikely*) that we did not include in the model. However, since
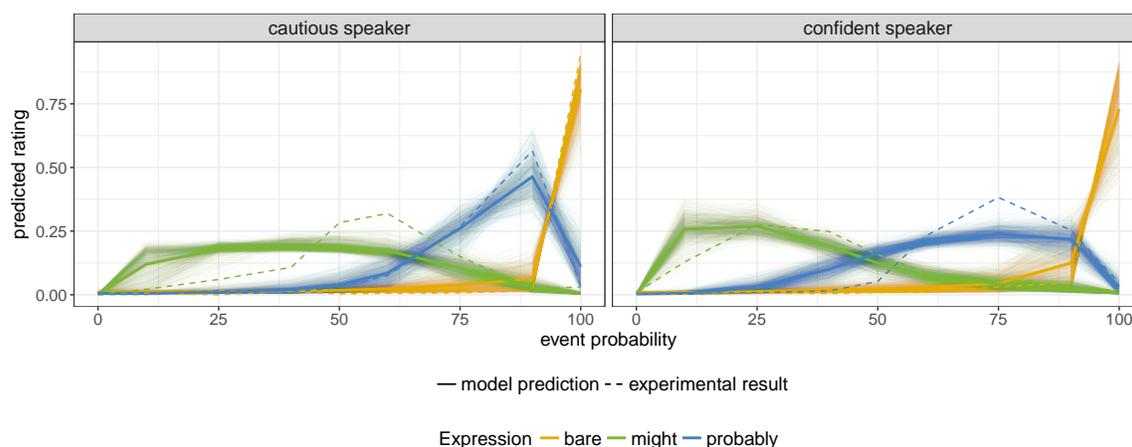
*Figure 16*. Predictions of *threshold distributions and costs* model and data from Experiment 3. The thin lines around the mean show the distribution of model predictions.

fine-tuning the set of alternative utterances would not change the qualitative predictions of the model and would not provide additional theoretical insights, we leave a more detailed exploration of this issue to future work.

## 7   General Discussion

While adaptation in language is a widely attested phenomenon, the exact nature of the representations that are updated during semantic/pragmatic adaptation has remained largely a mystery. In this paper we attempted to rectify this situation by doing three things: first, to investigate whether semantic/pragmatic adaptation occurs at all in a novel domain, that of uncertainty expressions; second, to investigate the nature of the representations that are updated during adaptation via the comparison of computational models of adaptation formulated within a Bayesian pragmatic modeling framework; and third, to test a novel prediction resulting from the application of this model to interpretation.

In two production expectation experiments (Experiments 2a and 2b), we found that listeners adapt to speakers who vary in their use of uncertainty expressions. This result confirms that the findings by Yildirim et al. (2016) also extend to the class of uncertainty expressions. In Experiment 3, we further found a novel effect of adaptation on utterance interpretation.

In a series of model comparisons, we found strong evidence for listeners updating their beliefs about both the speaker's lexicon as well as the speaker's preferences, which suggests that semantic/pragmatic adaptation is a result of updating both of these types of expectations. We further found that modeling the adaptation process as an instance of Bayesian belief updating explains participants' post-adaptation behavior in both the production expectation (Experiments 2a and 2b) and comprehension (Experiment 3) experiments.

We next discuss the implications of these results for other accounts of adaptation and for semantic theories of uncertainty expressions, as well as methodological implications. We then turn to limitations of the current results and account, as well as to fruitful future research avenues this work opens.

## 7.1 Implications for and relation to other accounts of adaptation

The model in this paper is formulated at the computational level (Anderson, 1990; Marr, 1982) and is therefore directly only comparable to other computational models. However, we can still assess the compatibility of our findings with mechanistic accounts. We first discuss the relation to existing computational models of adaptation and then discuss what the results tell us about existing mechanistic accounts of adaptation.

The model presented above follows several other computational models of linguistic adaptation that are based on Bayesian belief updating, including models of phonetic adaptation (Kleinschmidt & Jaeger, 2015), syntactic adaptation (Kleinschmidt et al., 2012), adaptation in the interpretation of prosodic cues (Roettger & Franke, 2019), and adaptation to variable use of the quantifiers *some* and *many* (Qing, 2014). Considering that these models that are all based on the same belief updating procedures can explain adaptation behavior from a range of linguistic domains, it is possible that linguistic adaptation is a result of cognitive processes that operate in a similar fashion at all levels of linguistic representation.

At the semantic and pragmatic level, Hawkins et al. (2017) proposed a model very similar to ours to explain the formation of conceptual pacts (Clark & Wilkes-Gibbs, 1986). Their model is based on the assumption that speakers and listeners have uncertainty about the lexicon (see also Bergen, Levy, & Goodman, 2016) and that in interaction, speakers and listeners update their beliefs about the lexicon akin the updating of threshold distributions in our model, which provides further evidence that belief updating plays an important role in interactive language processing.

In the space of mechanistic accounts, Pickering and Garrod (2004) argued that a lot of partner-specific linguistic behavior can be explained in terms of priming, i.e., the automatic activation of linguistic representations when a speaker produces an utterance or a listener hears an utterance. Their account has the appeal of explaining why partner-specific language use often appears to happen automatically and effortlessly. But without additional stipulations, their account is not compatible with the results from our experiments. In particular, such a priming account predicts that only the number of exposures should have an effect on language: with repeated exposure, the activation of the representations of lexical items like *might* and *probably* should increase, and therefore participants should be more likely to expect the speaker to produce these utterances. However, in Experiment 2b, we found evidence against such an account: in this experiment, participants in both conditions were exposed to the same number of each utterance, so according to a priming account, we should not find a difference between speaker bias conditions. Yet we did. In a more recent proposal, Pickering and Garrod (2013) agued that at least sometimes listeners perform *prediction-by-association* when processing an utterance, that is, listeners make predictions about what the speaker would say based on the context and their experience with the speaker. This appears to be compatible with our computational adaptation model but more details need to be worked out about how such predictions operate at the implementational level (Marr, 1982).

In a second line of work, Horton and Gerrig (2005, 2016) argued that partner-specific language use can be explained by an episodic memory account (Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2001). According to this account, individual linguistic events are

encoded together with speaker information and the world state in memory, which results in speaker-specific linguistic representations. This account is compatible with our findings, if we assume that individual utterance-world state pairs are stored in memory together with the speaker's identity, and that some additional inference mechanism gives rise to the more complex pragmatic behavior that we observed in our experiments.

## 7.2   Implications for the semantics of uncertainty expressions

Our results also have implications for semantic theories of uncertainty expressions. The finding that listeners rapidly update their beliefs about semantic thresholds of uncertainty expressions suggests that the semantics of these expressions is highly dynamic and context-sensitive. This is broadly compatible with recent theoretical accounts of probability operators (a subset of uncertainty expressions; e.g., Lassiter, 2017; Lassiter & Goodman, 2015; Yalcin, 2010), which state that the meanings of probability operators are highly dynamic and largely determined by the context. Our results suggest that the meaning of uncertainty expressions is even more dynamic than predicted by some of these accounts. First, we show that this dynamicity extends to a broader set of uncertainty expressions than is considered by some of these accounts (e.g., *might*; see also Lassiter, 2017, for arguments for all uncertainty expressions having a threshold semantics). Second, while these accounts generally assume that the main source of variability in interpretation is the probability of the event embedded under the uncertainty expressions, we find that knowledge of speaker identity also importantly contributes variability.

Dynamic and context-sensitive semantics have also been proposed for many other types of expressions. For example, Clark and Gerrig (1983) argued that speakers and listeners are able to compute novel senses of nouns and verbs on the fly. Similarly, in the domain of gradable adjectives such as *tall*, Kennedy (2007) and many others have argued that the interpretation of these adjectives crucially depends on contextual parameters. Considering the prevalence of dynamic meanings for so many other types of expression, it is therefore not surprising that the interpretation of uncertainty expressions also appears to be highly context-sensitive.

## 7.3   Methodological implications

Our results also have implications for conducting experiments. First, the finding that listeners adapt to the statistics of their environment within a short experiment suggests that experimenters should be cognizant of potential adaptation effects when probing production expectations or interpretations of uncertainty expressions (see also Jaeger, 2010).

Further, the results of Experiment 1, and in particular, the finding that participants' expectations about the use of utterances in the experiment strongly depended on the alternative utterances that we provided, highlights the need to be cautious about drawing general conclusions about expectations of use from single experiments. For example, had we only considered the results from the *bare-might* condition (see Figure 3), we might have concluded that "might" is an expected expression to communicate an event probability of 75%, whereas if we had only considered the results from the *might-probably* condition we might have instead concluded that it is *not* an expected expression to communicate an event probability of 75%. This is where explicit modeling of the sort we have engaged in here is

hugely helpful: formulating a concrete linking function which models the effects of alternatives allows for inferring the latent meanings of utterances by combining data from different experiments (see also Franke, 2014; Peloquin & Frank, 2016, for similar approaches).

### 7.4 Limitations and future directions

One potential limitation of the present research is that the paradigm is not fully interactive and that it does not involve any non-linguistic task, since participants only listened to pre-recorded utterances during the exposure phase. While this is clearly different from everyday dialog, it mirrors other everyday situations such as listening to someone talk on the radio or on TV. Further, since we only instructed participants to passively observe the interactions and the observations were not relevant for any non-linguistic task, we would expect that participants paid even more attention to the speaker's behavior if it had been relevant for a task. Investigating semantic/pragmatic adaptation in a less scripted setting is an important area for future research.

Throughout this paper, we made the assumption that listeners form *speaker-specific* production expectations. However, since all our experiments had a between-subjects design, it could be that participants were only adapting to the experimental situation, independent of the speaker. This seems unlikely given the results reported by Yildirim et al. (2016), who found that participants formed speaker-specific production expectations after being exposed to multiple speakers whose use of quantifiers differed. Moreover, (Schuster & Degen, 2019) have provided evidence of speaker-specific adaptation to uncertainty expressions. However, exactly which aspects of a situation (e.g., the speaker, the topic of conversation, the visual context, etc.) listeners adapt to is an issue that merits further investigation.

One advantage of formalizing a theory as a computational model is that the model makes concrete predictions to test in future experiments. For example, the proposed model is able to make quantitative predictions about the relation between the number of exposure trials and the size of the adaptation effect. Qualitatively, the model predicts that more exposure should lead to more adaptation, for which some evidence is reported by Schuster and Degen (2019). However, a systematic investigation of whether the model makes the correct quantitative predictions remains to be conducted.

Further, the presented adaptation model is built around the assumption that the utility of an utterance is exclusively determined by the informativeness and the cost of the utterance. However, it has been observed that other speaker goals such as being polite or convincing could also factor into the interpretation of uncertainty expressions (see e.g, Holtgraves & Perdew, 2016; Juanchich & Sirota, 2013; Pighin & Bonnefon, 2011). It could therefore be that, for example, listeners "explain away" the behavior of a *"confident"* speaker if the context suggests that the speaker has an incentive to be encouraging or has additional goals besides being informative (see also Yoon, Tessler, Goodman, & Frank, 2016, 2017). Investigating whether listeners draw such complex inferences could provide insight about which kind of potential speaker goals enter into listeners' pragmatic reasoning process.

### 7.5 Conclusion

We began with the puzzle of how to reconcile the assumption of stable utterance alternatives required for pragmatic reasoning with the rampant variability in speakers' language

use. The work reported here, building on much previous work on adaptation, suggests that this apparent tension is easily resolved if listeners form speaker-specific utterance expectations that they can recruit when encountering that same speaker again.

In a series of web-based experiments, we found that after exposure to a specific speaker, listeners rapidly update their expectations about which uncertainty expressions that speaker is likely to produce to describe varying event probabilities, and that these updated expectations also transferred to updated interpretations. We provided a formal account of semantic/pragmatic adaptation and modeled this behavior using a Bayesian cognitive model which assumes that (listeners reason about) speakers (who) efficiently trade off utterance informativeness and cost. Through a series of simulations we found strong evidence for semantic/pragmatic adaptation being a result of updated beliefs about a specific speaker's meaning of uncertainty expressions as well as the speaker's utterance preferences.

## References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. doi:10.1121/1.1528172

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Lawrence Erlbaum Associates.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189. doi:10.1016/j.wocn.2011.09.001

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*(20), 1–83. doi:10.3765/sp.9.20

Berry, D. (2004). *Risk, communication, and health psychology*. London: Open University Press.

Brennan, S. E. & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. doi:10.1037/0278-7393.22.6.1482

Brennan, S. E. & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291. doi:10.1111/j.1756-8765.2009.01019.x

Clark, H. H. (1990). Quantifying probabilistic expressions: Comment. *Statistical Science*, *5*(1), 12–16. doi:doi:10.1214/ss/1177012242

Clark, H. H. & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 591–608. doi:10.1016/S0022-5371(83)90364-X

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi:10.1016/0010-0277(86)90010-7

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. doi:10.1016/j.cognition.2008.04.004

Degen, J. & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172–201. doi:10.1111/cogs.12227

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20. doi:10.1016/j.cognition.2019.01.001

Doupnik, T. S. & Richter, M. (2003). Interpretation of uncertainty expressions: A cross-national study. *Accounting, Organizations and Society*, *28*, 15–35.

Fine, A. B. & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(9), 1362–1376. doi:10.1037/xlm0000236

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10), 1–18. doi:10.1371/journal.pone.0077661

Finegan, E. & Biber, D. (2001). Register variation and social dialect variation: The Register Axiom. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 235–267). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511613258.015

Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998. doi:10.1126/science.1218633

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the 36th annual conference of the Cognitive Science Society (CogSci 2014)*.

Franke, M. & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44. doi:10.1515/zfs-2016-0002

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton: CRC Press.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *0*, 1–7. doi:10.1080/00031305.2018.1549100

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. doi:10.1037/0033-295X.105.2.251

Goodman, N. D. & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi:10.1016/j.tics.2016.08.005

Goodman, N. D. & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184. doi:10.1111/tops.12007

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). New York: Academic Press.

Grodner, D. & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (Eds.), *The processing and acquisition of reference* (Chap. 10, pp. 239–272). Cambridge: MIT Press.

Hacquard, V. (2011). Modality. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning* (Chap. 58, pp. 1484–1515). Berlin: Mouton de Gruyter.

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. doi:10.3758/s13421-018-0808-6

Harrington, J., Palethorpe, S., & Watson, C. (2000). Monophthongal vowel changes in received pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association*, *30*(1-2), 63–78. doi:10.1017/S0025100300006666

Hawkins, R. X., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th annual meeting of the Cognitive Science Society (CogSci 2017)* (pp. 482–487).

Herbstritt, M. & Franke, M. (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71. doi:10.1016/j.cognition.2018.11.013

Holtgraves, T. & Perdew, A. (2016). Politeness and the communication of uncertainty. *Cognition*, *154*, 1–10. doi:10.1016/j.cognition.2016.05.005

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–4). Washington, D.C.: Georgetown University Press.

Horton, W. S. & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. doi:10.1016/j.cognition.2004.07.001

Horton, W. S. & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science*, *8*(4), 780–795. doi:10.1111/tops.12216

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. doi:10.1016/j.cogpsych.2010.02.002

Johnson, K. (1997). Speaker perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Francisco: Morgan Kaufmann.

Juanchich, M. & Sirota, M. (2013). Do people really say it is "likely" when they believe it is only "possible"? Effect of politeness on risk communication. *Quarterly Journal of Experimental Psychology*, *66*(7), 1268–1275. doi:10.1080/17470218.2013.804582

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. doi:10.1016/j.cognition.2012.03.001

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. doi:10.1007/s10988-006-9008-0

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the 34th annual meeting of the Cognitive Science Society (CogSci 2012)* (pp. 599–604).

Kleinschmidt, D. F. & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi:10.1037/a0038695

Kraljic, T. & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. doi:10.1016/j.cogpsych.2005.05.001

Kraljic, T. & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. doi:10.1016/j.jml.2006.07.010

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantik: Ein internationales handbuch zeitgenössischer forschung.* (Chap. 29, pp. 639–650). Berlin: Mouton de Gruyter.

Kroczek, L. O. H. & Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Scientific Reports*, *7*(17581), 1–9. doi:10.1038/s41598-017-17907-9

Kruschke, J. (2015). *Doing Bayesian data analysis* (2nd). London: Academic Press.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342. doi:10.1016/j.cognition.2014.05.017

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In *Proceedings of the 34th annual meeting of the Cognitive Science Society (CogSci 2012)* (pp. 647–652).

Lassiter, D. (2017). *Graded modality: Qualitative and quantitative perspectives.* Oxford: Oxford University Press.

Lassiter, D. & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, *136*, 123–134. doi:10.1016/j.cognition.2014.10.016

Lassiter, D. & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836. doi:10.1007/s11229-015-0786-1

Lewis, D. (1969). *Convention: A philosophical study.* Cambridge: Harvard University Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. doi:10.1037/h0020279

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*(5), 696–713. doi:10.1177/0272989X07307271

Liu, L., Burchill, Z., Tanenhaus, M. K., & Jaeger, T. F. (2017). Failure to replicate talker-specific syntactic adaptation. In *Proceedings of the 39th annual meeting of the Cognitive Science Society (CogSci 2017)* (pp. 2616–2621).

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Oxford: John Wiley.

Marr, D. (1982). *Vision.* San Francisco: W.H. Freeman and Company.

Metzing, C. & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213. doi:10.1016/S0749-596X(03)00028-7

Moss, S. (2018). *Probabilistic knowledge.* Oxford: Oxford University Press. doi:10.1093/oso/9780198792154.001.0001

Myslín, M. & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition*, *147*, 29–56. doi:10.1016/j.cognition.2015.10.021

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peloquin, B. & Frank, M. C. (2016). Determining the alternatives for scalar implicature. In *Proceedings of the 38th annual conference of the Cognitive Science Society (CogSci 2016)*.

Pepper, S. & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*(1), 95–101. doi:10.1016/0092-6566(74)90049-X

Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–226. doi:10.1017/S0140525X04000056

Pickering, M. J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. doi:10.1017/S0140525X12001495

Pierrehumbert, J. B. (2001). Exemplar dynamics. In P. J. Hopper & J. L. Bybee (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–158). Amsterdam: John Benjamins Publishing Company. doi:10.1075/tsl.45.08pie

Pighin, S. & Bonnefon, J.-F. (2011). Facework and uncertain reasoning in health communication. *Patient Education and Counseling*, *85*(2), 169–172. doi:10.1016/j.pec.2010.09.005

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 1–18. doi:10.3389/fpsyg.2015.02035

Pogue, A. & Tanenhaus, M. K. (2018). Learning from uncertainty: Exploring and manipulating the role of uncertainty on expression production and interpretation. In *Proceedings of the 40th annual meeting of the Cognitive Science Society (CogSci 2018)* (pp. 2266–2271).

Politi, M. C., Han, P. K. J., & Col, N. F. (2007). Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, *27*(5), 681–695. doi:10.1177/0272989X07307270

Qing, C. (2014). *Quantiative social-cognitive experimental pragmatics* (Master's thesis, Universiteit van Amsterdam).

Qing, C. & Franke, M. (2015). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Proceedings of the 24th semantics and linguistic theory conference (SALT 24)* (pp. 23–41). doi:10.3765/salt.v24i0.2412

Roettger, T. B. & Franke, M. (2019). Evidential strength of intonational cues and rationaladaptation to (un-)reliable intonation. *Cognitive Science*, *43*, e12745. doi:10.1111/cogs.12745

Schuster, S. & Degen, J. (2019). Speaker-specific adaptation to variable use of uncertainty expressions. In *Proceedings of the 41st annual meeting of the Cognitive Science Society (CogSci 2019)*.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Swanson, E. (2008). Modality in language. *Philosophy Compass*, *3*(6), 1193–1207. doi:10.1111/j.1747-9991.2008.00177.x

Tessler, M. H. & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395–436. doi:10.1037/rev0000142

van der Walt, S., Colbert, S. ., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering*, *13*(2), 22–30. doi:10.1109/MCSE.2011.37

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348–365. doi:10.1037//0096-3445.115.4.348

Weiner, E. J. & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(1), 29–58. doi:10.1017/S0022226700007441

Yalcin, S. (2010). Probability operators. *Philosophy Compass*, *5*(11), 916–937. doi:10.1111/j.1747-9991.2010.00360.x

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi:10.1016/j.jml.2015.08.003

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual conference of the Cognitive Science Society (CogSci 2016)* (pp. 2771–2776).

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the 39th annual conference of the Cognitive Science Society (CogSci 2017)*.

**Appendix**

**A Effect of color in Experiment 1**

As mentioned in a footnote, we ran the norming studies in three batches using three slightly different procedures across conditions. We originally ran condition 0 (*bare-might*) as a pilot condition. In the results, we noted that participants did not differ in their ratings depending on whether the girl asked for a blue or an orange gumball ($R^2(27) = 0.997$ between mean ratings for blue and orange trials). To lower the number of trials, we therefore asked each participant to provide ratings for only one of the two colors (randomized across participants) for the next batch of conditions (conditions 1-14). We found that in some conditions, this led to small differences in ratings between participants who always rated utterances with *blue* and participants who always rated utterances with *orange* ($R^2(27)$ between 0.864 and 0.984). We hypothesize that this is a result of participants paying less attention if they were asked to do exactly the same task over and over again (in condition 0, the color and the associated utterances could change across trials). In order to verify the stability of our results, we replicated one of the conditions, condition 5 (*might-probably*), and had participants provide two ratings for each color and gumball proportion. We found that despite the lower correlation between average ratings for utterances with *blue* and utterances with *orange* in the original run ($R^2(27) = 0.929$), there was a very high correlation between the average ratings independent of the color of the original study and the average ratings of the replication ($R^2(27) = 0.975$), which suggests that the average ratings largely do not depend on whether we ask participants to provide ratings for both colors or just one color. Nevertheless, we used the modified procedure in which we asked participants to provide 2 ratings for each color and gumball proportion for the last batch of conditions (conditions 15-20). In all conditions in which we asked people to provide ratings for utterances with both colors, the correlation between average ratings for utterances with *blue* and utterances with *orange* was almost perfect ($R^2(27) > 0.988$).

## B Additional results of Eperiment 1.

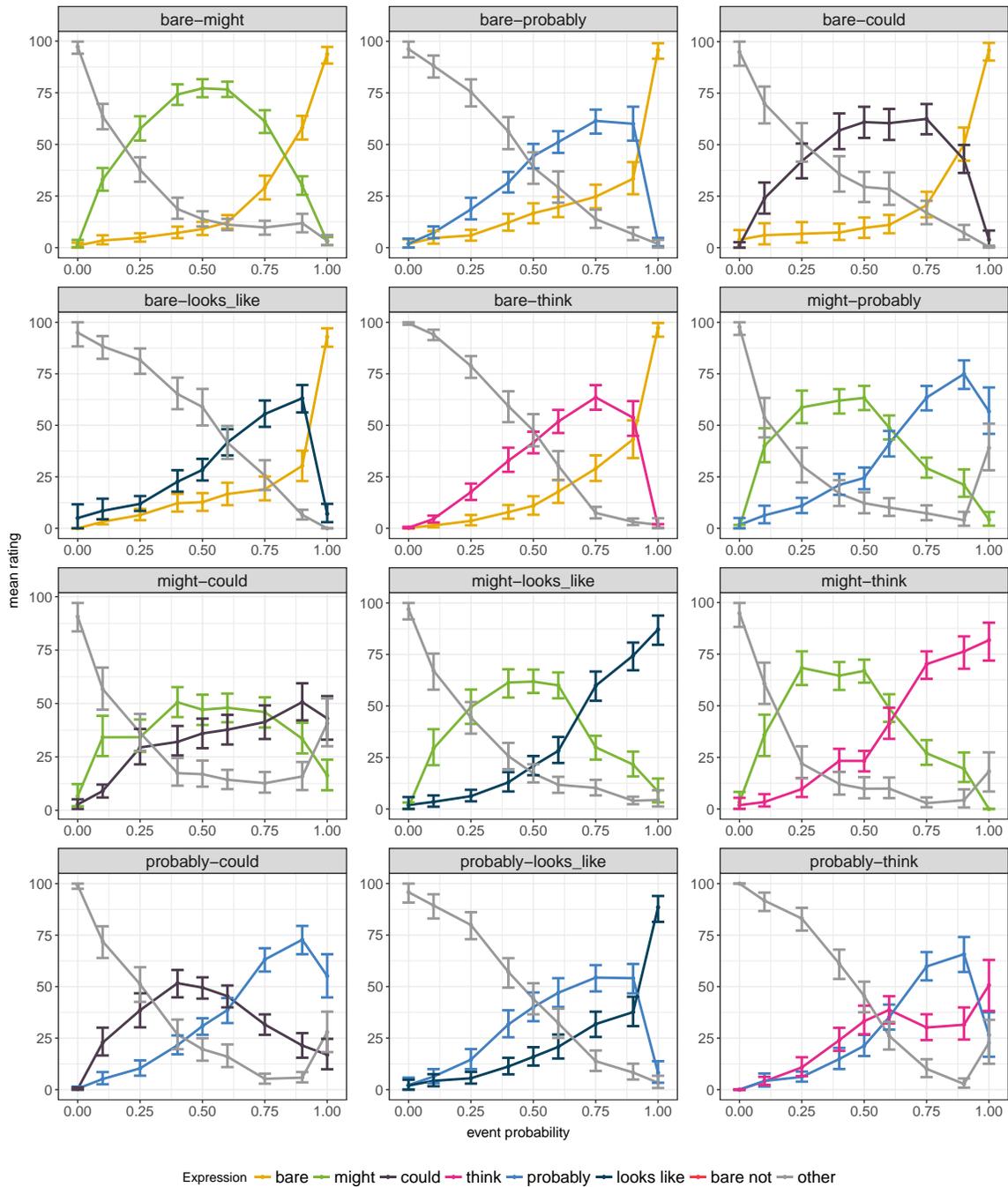Figures 17 and Figures 18 show the results from all conditions in Experiment 1.



*Figure 17*. Results of Experiment 1 – Part 1. Error bars correspond to bootstrapped 95%-confidence intervals.
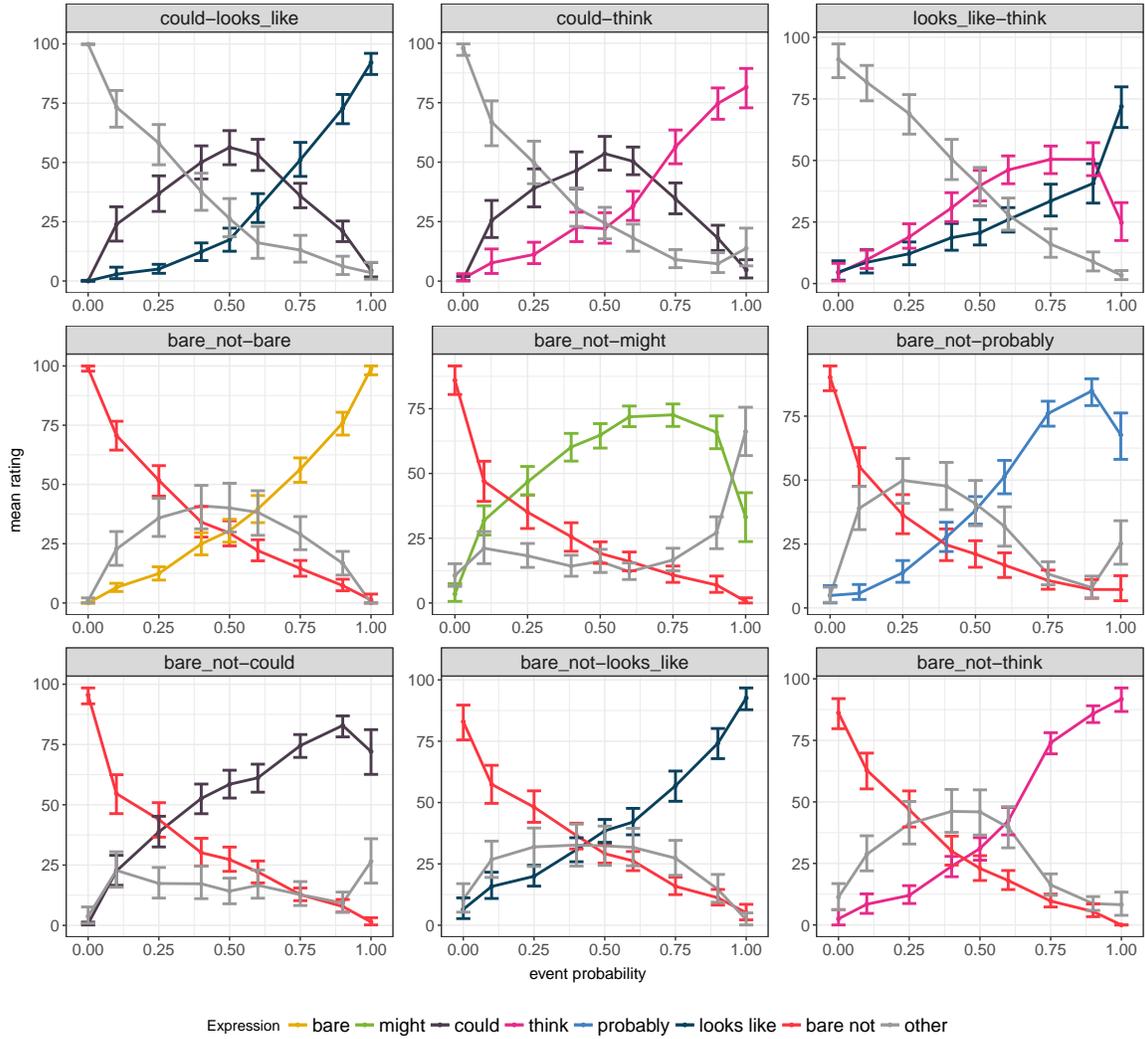
*Figure 18*. Results of Experiment 1 – Part 2. Error bars correspond to bootstrapped 95%-confidence intervals.

## C Model implementation details

The model presented above poses some challenges for performing Bayesian data analysis with considerable amounts of data. Concretely, the integral over threshold distributions

in the expected pragmatic speaker model $ES_1$ (repeated here) makes it hard to compute the distribution $ES_1$ given a set of parameters $\Theta$.

$$ES_1\left(u_e \mid \phi\right) = \int P(c) \int_0^1 P(\theta)S_1\left(u_e \mid \phi, \theta, c\right) d\theta \ dc$$

The reason for this is two-fold: First, there is no analytical solution for this integral, and second, since $S_1$ depends on thresholds for all uncertainty expressions $P(\Theta)$ is a multidimensional distribution which cannot be easily approximated.

We solve this issue by introducing two approximations. First, we discretize the threshold distributions by distributing the probability mass of the Beta distributions across 20 equally-wide bins, resulting in a discrete probability distribution $P_d(\theta)$ (see Tessler and Goodman, 2019 for a similar approach). Since all event probabilities for which participants had to provide ratings in the experiments were multiples of 5%, we do not lose any accuracy and gain the advantage that we can now sum over a discrete probability space:[16]

$$ES_1\left(u_e \mid \phi\right) = \sum_\theta P_d(\theta)S_1\left(u_e \mid \phi, \theta, c\right)$$

While this approximation can in theory be computed exactly, its computation remains intractable even for the small number of utterances that we included in our model. Note that the discrete version of the vector of thresholds $\theta$ has one dimension with 20 possible values for each utterance, which implies there are $20^{|U|}$ possible assignments of $\theta$. This means for estimating parameters for a model with 7 utterances, we would have to sum over $20^7 = 1.28 \times 10^9$ parameterizations of the pragmatic speaker model $S_1$ to compute the likelihood for one sample of parameters in the BDA.

We solve this problem through another approximation, which exploits the fact that $S_1(u_e \mid \phi, \theta, c)$ only depends on the thresholds for uncertainty expressions other than $e$ for the normalization term. We approximate the normalization term by marginalizing over $\theta'_e$ and thus making $S'_1$ independent of all thresholds except $\theta_e$:

$$\widetilde{S_1}(u_e \mid \phi, \theta_e, c) = \frac{exp \ \mathbb{U}(\phi, u_e, \theta_e, c)}{exp \ \mathbb{U}(\phi, u_e, \theta_e, c) + \sum_{u'_e \neq u_e} \sum_{\theta_{e'}} P_d(\theta_{e'}) \ exp \ \mathbb{U}(\phi, u_{e'}, \theta_{e'}, c)},$$

where $\mathbb{U}(\phi, u_e, \theta_e, c) = \log L_0(\phi \mid u_e, \theta_e) - c(u)$ is the speaker utility as defined in the main text.

This approximation allows us to define the following approximation of $ES_1$, which is tractable since we only have to sum over all values of one threshold instead of all combinations of thresholds:

$$\widetilde{ES_1}(u_e \mid \phi) \propto \sum_{\theta_e} P_d(\theta_e)\widetilde{S_1}\left(u_e \mid \phi, \theta_e, c\right)$$

This approximation leads to identical results as $ES_1$ if each threshold distributions assigns all probability mass to one value, i.e., if we have point estimates for thresholds. To

---

[16]In our data analysis procedure, we assumed that the distribution over cost functions, $P(c)$, is a delta distribution which assigns all probability mass to the condition-specific cost function $c(u, \mathscr{C})$ parameterized by the cost parameter $\gamma$. Since this implies that $P(c)$ is zero for all other cost functions, we can omit the integral and replace $c$ with the condition-specific cost function, which we implicitly did here.
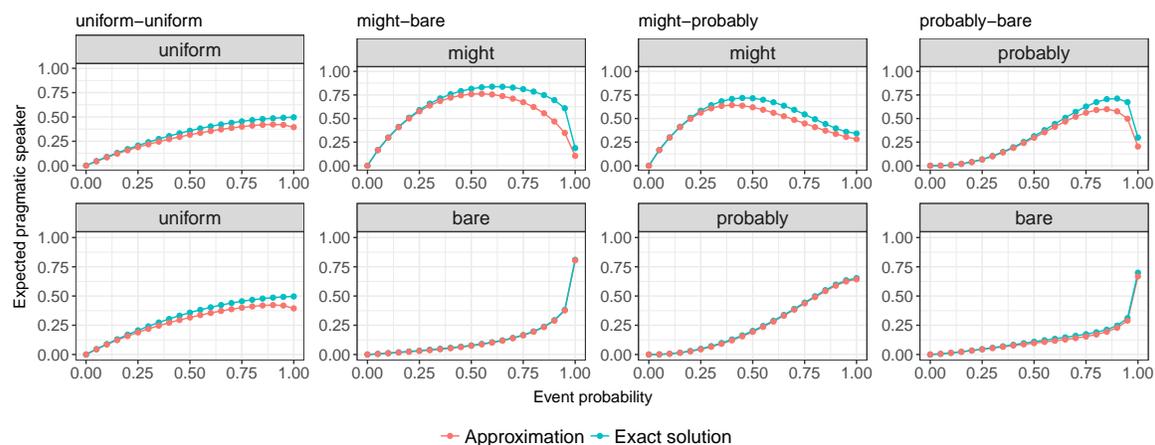
*Figure 19*. Predictions of exact and approximate expected pragmatic speaker model for different combinations of thresholds. The leftmost panels (uniform) shows predictions of both models if both utterances have uniform threshold distributions, i.e., threshold distributions with very high variance. The other panels show model predictions under the assumption that the utterances have the threshold distributions that we inferred in Section 3.

assess how much $ES_1$ and its approximation, $\widetilde{ES_1}$ deviate when the threshold distributions have non-zero variance, we performed several simulations, with different threshold distributions. For these simulations, we assume that there are only two possible utterances, which makes the computation of $ES_1$ tractable.

Figure 19 shows the results of these simulations. As these plots show, the approximate model $\widetilde{ES_1}$ is a very close approximation of the expected pragmatic speaker model $ES_1$, which suggests that this approximation should only minimally affect our modeling results.

The model is implemented in Python using the scikit-learn (Pedregosa et al., 2011) and numpy (van der Walt, Colbert, & Varoquaux, 2011) libraries.

## D  Additional model predictions

Figures 20 and Figures 21 show the model predictions and the results from all conditions in Experiment 1.
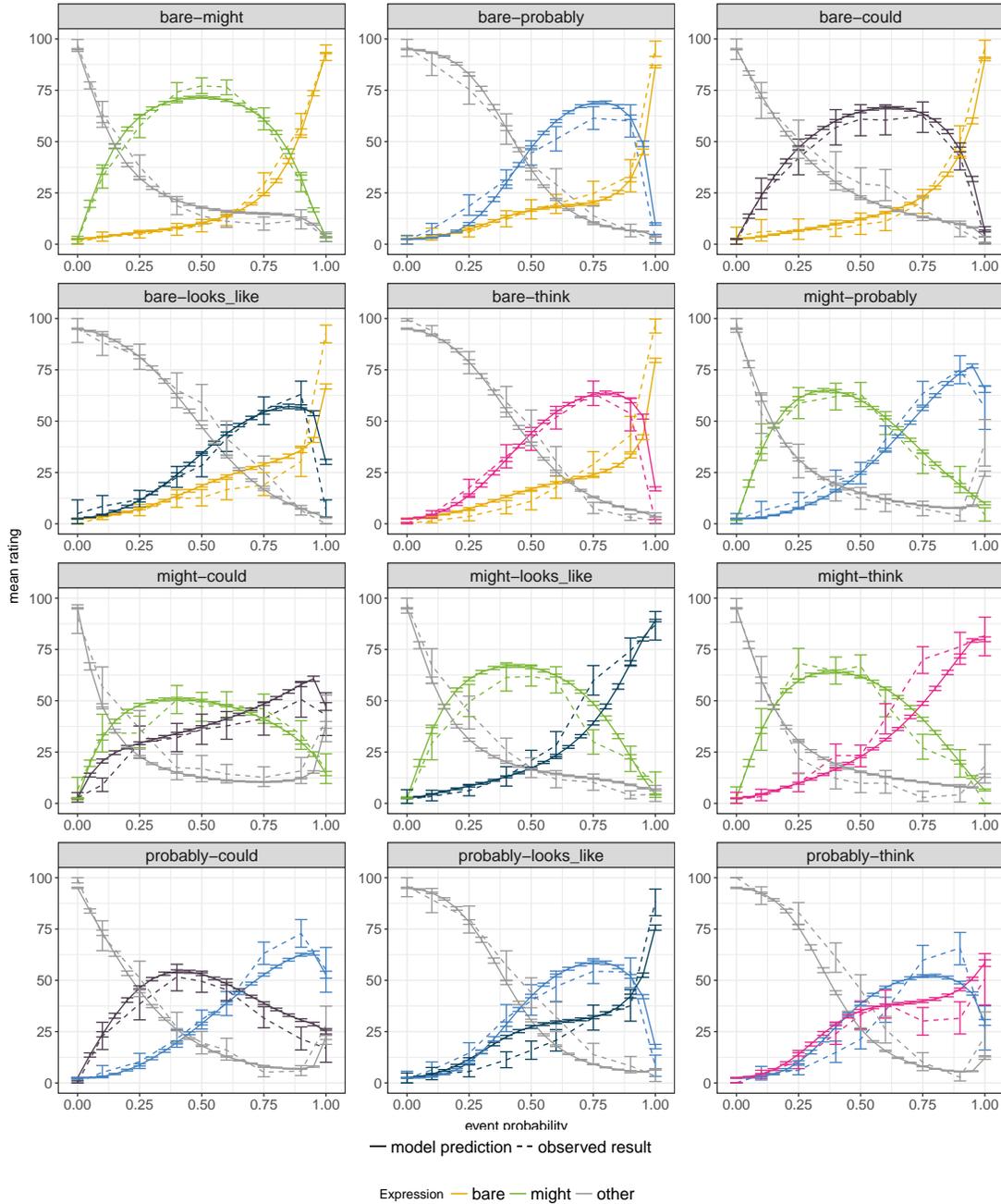


*Figure 20*. Model predictions and results of Experiment 1 – Part 1. Error bars correspond to 95% high density intervals (model predictions) and bootstrapped 95%-confidence intervals (observed results).
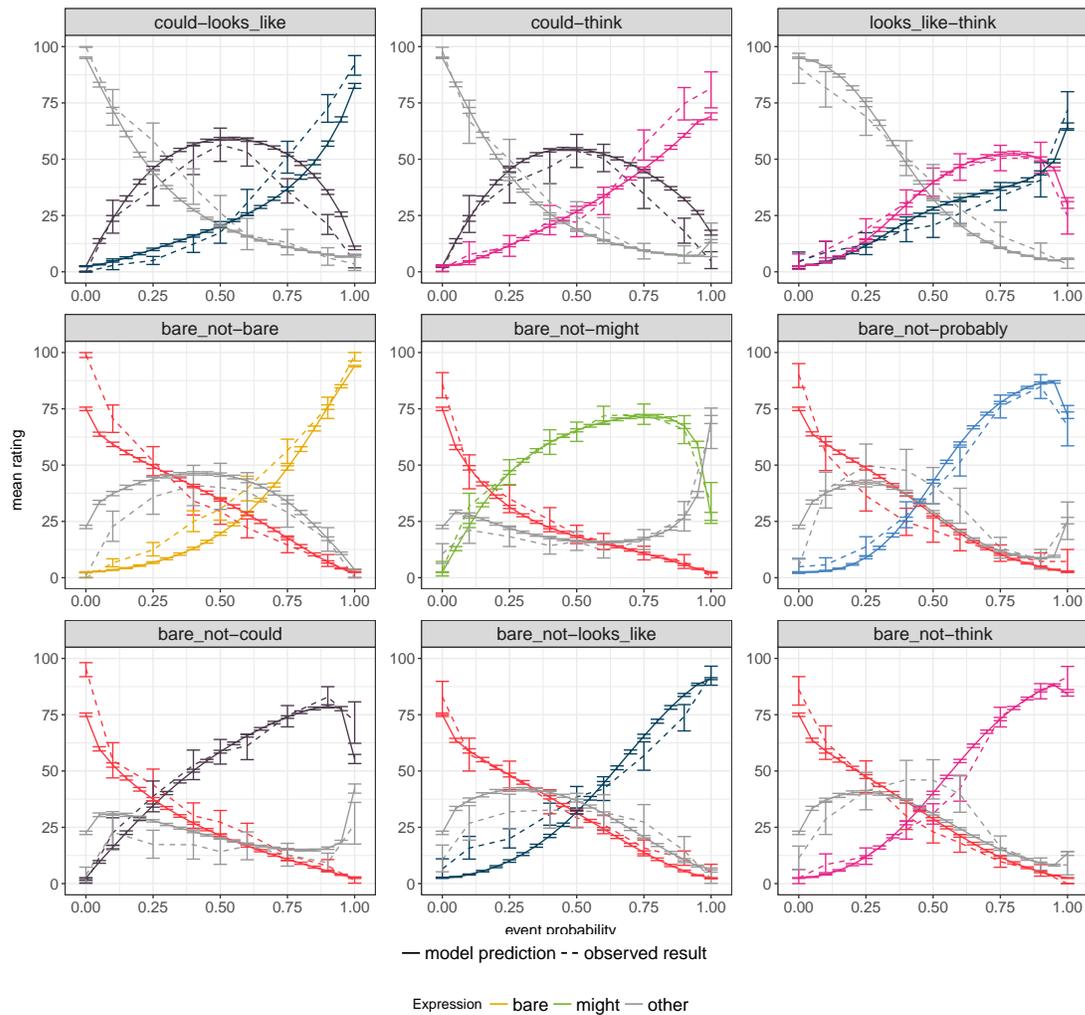
*Figure 21.* Model predictions and results of Experiment 1 – Part 2. Error bars correspond to 95% high density intervals (model predictions) and bootstrapped 95%-confidence intervals (observed results).

## E Model simulations for Experiment 2b

Figures 22, 23, and 24 show the posterior predictions of the model simulations for Experiment 2b, the post-adaptation threshold distributions, and the post-adaptation costs, respectively.
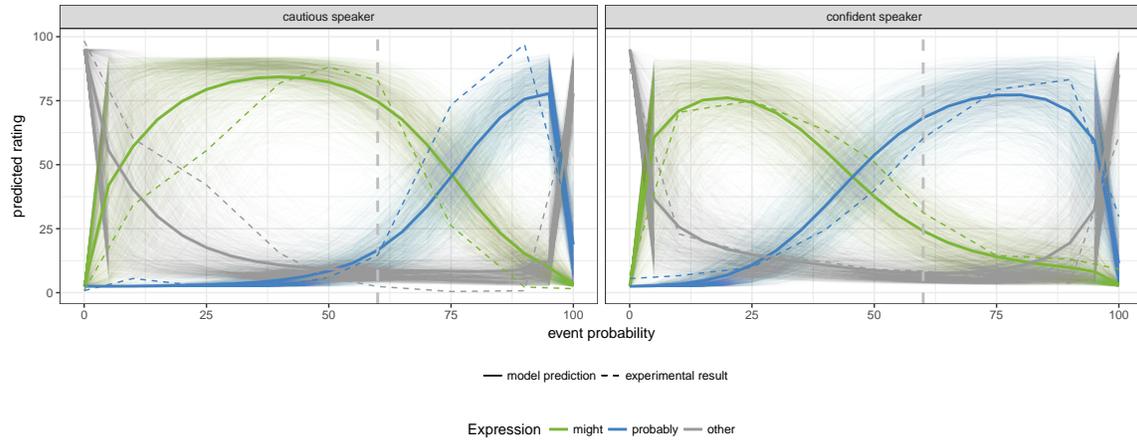
*Figure 22.* Post-adaptation model predictions from simulations for Experiment 2b and experimental results. The solid lines shows the mean model predictions and the thin lines around the mean show the distribution of model predictions.
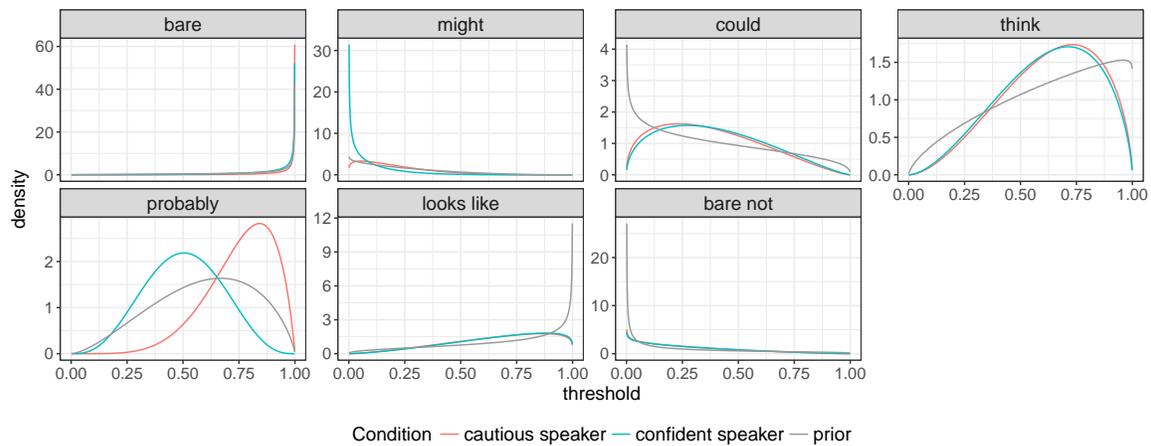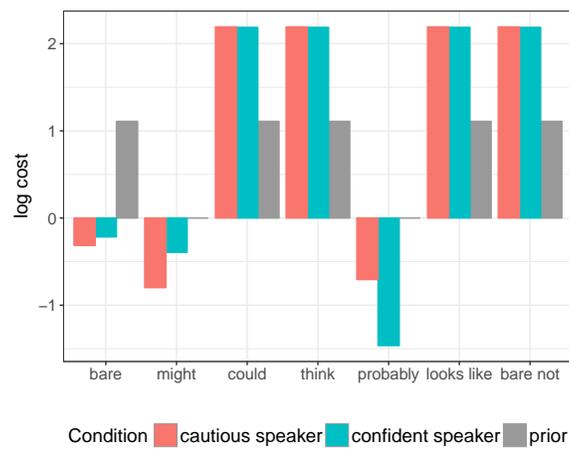


*Figure 23.* Post-adaptation threshold distributions from the simulations for Experiment 2b.

*Figure 24*. Post-adaptation *log* cost values from simulations for Experiment 2b. Note that the cost of MIGHT and PROBABLY in the norming data model was 1 and therefore the *log* cost for these utterances is 0.

### F Original interpretation experiment

As we mentioned in the main text, we originally ran a slightly different version of the comprehension experiment in which participants used sliders to rate which gumball machine they thought the speaker was describing. While the results were qualitatitvely the same as in the experiment reported in the main body of the paper, the use of sliders seemed to confuse some participants (see details below) and therefore we changed the procedure such that participants provided ratings by distributing coins. For the sake of completeness, we report the procedure and the results of the original experiment here.

**Participants.** We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2 which amounted to an hourly wage of approximately $10–$12. None of the participants had participated in any of the previous experiments.

**Materials and Procedure.** The exposure phase was identical as in the other adaptation experiments: participants were either exposed to a *cautious* speaker or a *confident* speaker. Six of the exposure trials included attention checks in which participants had to indicate whether they saw a grey X on the previous trial or not.

Similar to Experiment 3, the test trials probed participants' interpretations of the utterances MIGHT, PROBABLY, and BARE. On test trials, participants listened to a recording of the speaker they encountered during the exposure phase and then rated how likely they thought it was that the speaker saw different gumball machines. On each trial, like in Experiment 3, participants provided ratings for 9 gumball machines. However, unlike in Experiment 3, participants indicated their ratings by adjusting 9 sliders. Participants completed 6 test trials in total – one for each expression-color pair.

**Exclusions.** We excluded participants who failed more than 2 out of 6 attention checks, which led to 2 exclusions in the *cautious speaker* condition and 1 exclusion in the *confident speaker* condition.

**Analysis and Predictions.** As for Experiment 3, we expected that listeners interpret a more confident speaker's utterance to communicate a lower event probability than a more cautious speaker's utterance. We measured the interpretation of utterances by normalizing the ratings across the 9 gumball machines so that they sum to 1 and then computing the expected value for the proportion of blue and orange gumballs. We predicted that the expected values of target color gumball proportions after hearing MIGHT and PROBABLY were going to be larger in the *cautious speaker* condition than in the *confident speaker* condition.

**Results and Discussion.** Figure 25 shows the aggregated and normalized ratings for the two conditions. As predicted, participants provided higher ratings for gumball machines with higher target color percentages after hearing MIGHT and PROBABLY in the *cautious speaker* condition than in the *cautious speaker* condition. This also led to a significantly higher expected value for MIGHT ($t(75) = 3.05$, $p < 0.01$) and PROBABLY ($t(75) = 3.08$, $p < 0.01$) in the *cautious speaker* condition as compared to the *confident speaker* condition.

This means that qualitatively, the results are the same as in Experiment 3. However, since participants had the option to assign high ratings to all gumball machines (they could assign a maximum rating to each gumball machine if they wanted to), we noticed that
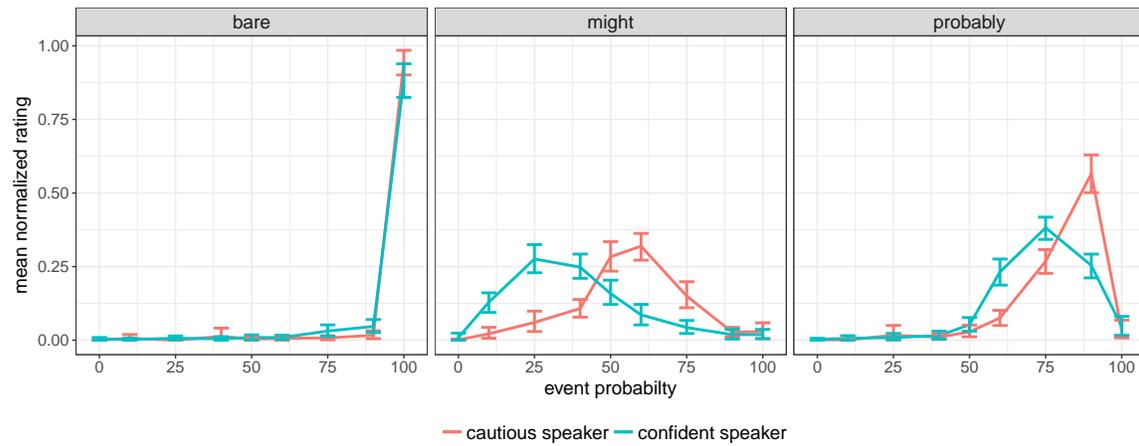
*Figure 25*. Aggregated post-exposure ratings from the original interpretation experiment.

many participants assigned very high ratings to most gumball machines and therefore did not indicate their interpretation of the utterance. Further, it seemed that some participants understood the instructions as rating the likelihood of getting a target color gumball and provided ratings proportional to the target color gumball proportion independent of the utterance. For these reasons, we revised the original paradigm as described in the main text and asked participants to indicate their interpretation using a limited set of coins, which appeared to be less confusing for participants.