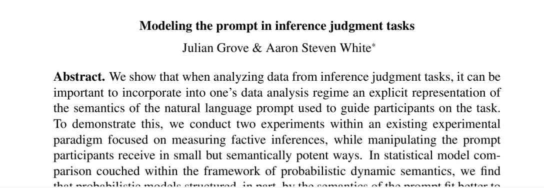
Modeling the prompt in inference judgment tasks

Julian Grove and Aaron Steven White

ELM 3 (University of Pennsylvania), June 13, 2024

FACTS.lab, Linguistics Department, University of Rochester

Joint work



Modeling the prompt in inference judgment tasks

Julian Grove & Aaron Steven White*

Abstract. We show that when analyzing data from inference judgment tasks, it can be important to incorporate into one's data analysis regime an explicit representation of the semantics of the natural language prompt used to guide participants on the task. To demonstrate this, we conduct two experiments within an existing experimental paradigm focused on measuring factive inferences, while manipulating the prompt participants receive in small but semantically potent ways. In statistical model comparison couched within the framework of probabilistic dynamic semantics, we find

Aaron Steven White

University of Rochester

Motivation

Common practice

Inference judgments in formal experiments:

- Some target linguistic expression, along with a context.
- A natural language prompt.
- A response instrument; e.g., a Likert scale, a slider scale, etc.

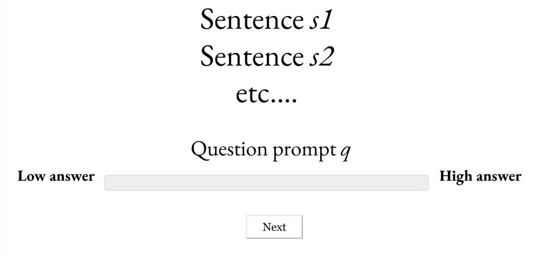
Models of inference data generally encode only representations of:

- the target expression plus context (via, e.g., model parameters).
- the response instrument (via, e.g., a link function or likelihood).

Today, we focus on the *prompt*.

What we're advocating

We should think of an experimental trial as a little discourse.



We model this discourse using *probabilistic* dynamic semantics.

- Sentences: Start with a prior distribution over *discourse states*. Update this prior with $\llbracket s1 \rrbracket$, then $\llbracket s2 \rrbracket$, etc.
- Question: Push $\llbracket q \rrbracket$ onto the QUD stack (Farkas and Bruce 2010; Roberts 2012).
- Answer: Pop $\llbracket q \rrbracket$ off the QUD stack; respond.

Upshot: probabilistic models of data and semantic analyses are one and the same.

Case study: factive predicates

Lots of recent experimental work on factive inferences.
See, e.g., Degen and Tonhauser (2021, 2022), Djärv and Bacovcin (2017), Djärv, Zehr, and Schwarz (2018), and Grove and White (2024).

(1) Jo loves that Mo Left.
$\rightsquigarrow$ Mo left.

Good case study because:

- Factivity is a rich discourse phenomenon with a nonetheless clear inferential profile.
- Factive inferences, in aggregate, display substantial *gradience*—a tricky phenomenon to analyze statistically.

Plan

- Illustrate the gradience exhibited by factive inferences in formal experiments.
- Show that we can improve models of this gradience by carefully representing the compositional semantics of the prompt in models of inference judgment data.
- Along the way, we illustrate data from two novel experiments which vary the prompt in subtle, but semantically potent ways.

Gradience in inference experiments

Gradience

What sorts of inference patterns arise from uses of factive predicates in an experimental setting?

- E.g., if you ask someone to rate the likelihood that Mo left, given that *Jo loves that Mo left* is true.

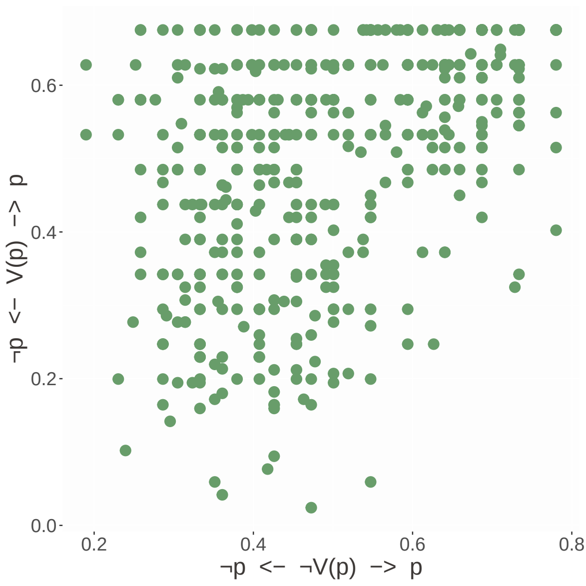
White and Rawlins (2018)

‘Someone {discovered, didn’t discover} that a particular thing happened.’

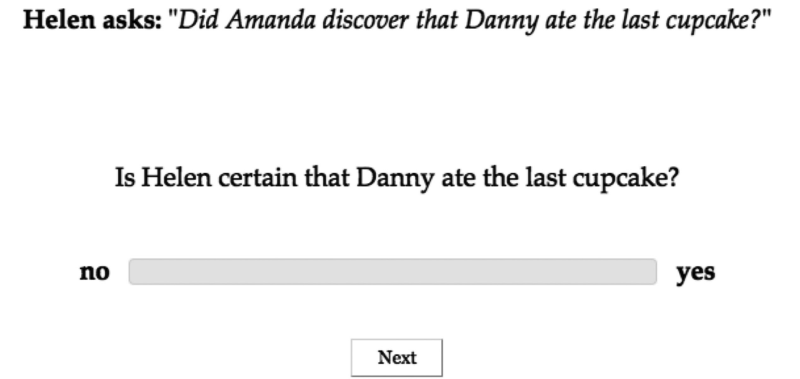
‘Did that thing happen?’

(*yes*, *maybe or maybe not*, *no*)

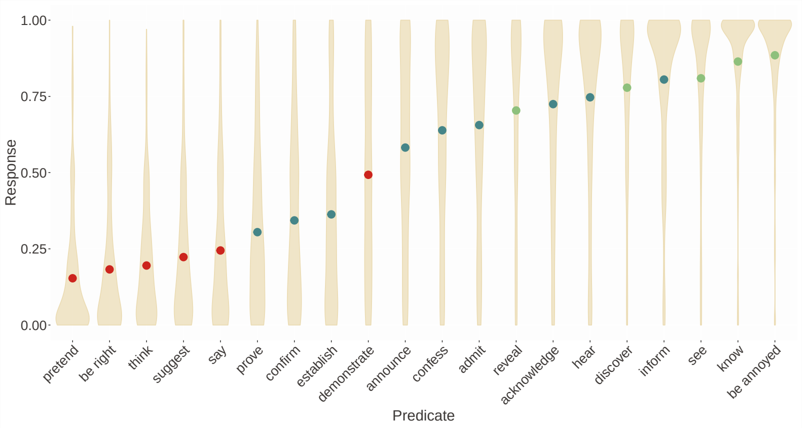
White and Rawlins (2018)



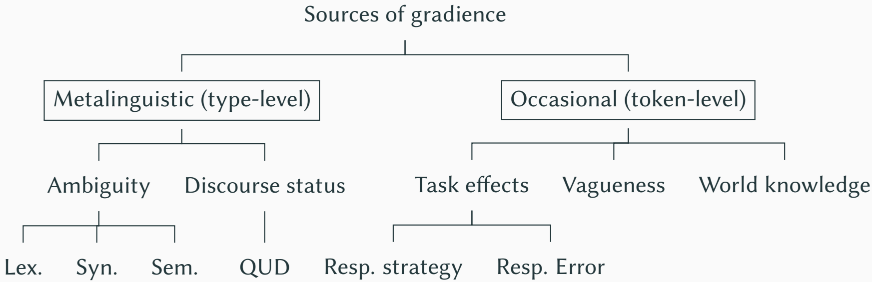
Degen and Tonhauser (2022)



Degen and Tonhauser (2022)



Possible sources of gradience



- Ambiguity: *run* (organizational) vs. *run* (locomotive)
- Vagueness: *X is tall* $\rightsquigarrow$ vagueness about X's height

In probabilistic dynamic semantics, we can formalize this distinction.

Modeling prompts dynamically

Main questions

- How do people represent their knowledge of factivity: is its gradience metalinguistic or occasional?

Factivity, presupposition projection, and the role of discrete knowledge in gradient inference judgments*

Julian Grove and Aaron Steven White
University of Rochester

Abstract We investigate whether the factive presuppositions associated with some clause-embedding predicates are fundamentally discrete in nature—as classically assumed—or fundamentally gradient—as recently proposed (Tonhauser, Beaver, and Degen 2018). To carry

  Grove and White (2024): gradience in factivity is like ambiguity, *not* vagueness.

- How should we capture the fine-grained semantics of the prompt used in eliciting judgments?
  - How does manipulating and modeling the prompt affect our earlier findings?

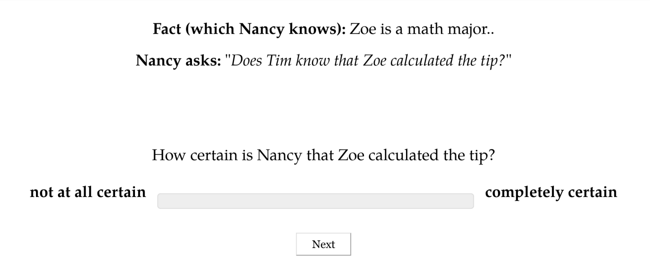
Experiments

Two experiments, differing only by the prompt used.
Following the paradigm of Degen and Tonhauser (2021).

Fact (which Nancy knows): Zoe is 5 years old. / Zoe is a math major.
Nancy asks: "Does Tim know that Zoe calculated the tip?"

- Experiment 1: How certain is Nancy that Zoe calculated the tip?
- Experiment 2: How likely is it that Nancy is certain that Zoe calculated the tip?

Experiment 1: the “how certain” task


Fact (which Nancy knows): Zoe is a math major..

Nancy asks: "*Does Tim know that Zoe calculated the tip?*"

How certain is Nancy that Zoe calculated the tip?

not at all certain **completely certain**

Next

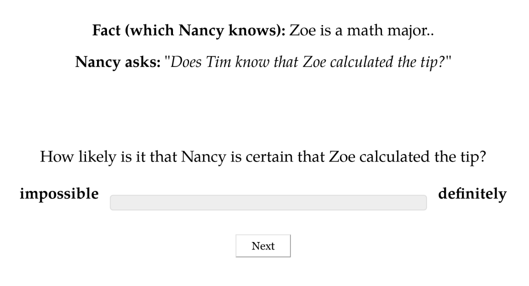
- Start with a prior distribution over *discourse states.*
  Update with ⟦*Zoe is a math major*⟧.
  Update with ⟦*Tim knows that Zoe calculated the tip*⟧.
- Push ⟦*How certain is Nancy that Zoe calculated the tip?*⟧ onto the QUD stack.
- Pop it off the QUD stack; respond with maximally informative answer.

Semantics of '*how certain is X that p*'

- Assumption: while *likely* predicates of degrees on a probability scale, *certain* predicates of degrees on a confidence scale (Klecha 2012).
- In practice, the scale associated with *certain* is truncated at the lower end, relative to the scale for *likely*.

Ask about details in the Q&A!

Experiment 2: the “how likely … certain” task


Fact (which Nancy knows): Zoe is a math major..

Nancy asks: "*Does Tim know that Zoe calculated the tip?*"

How likely is it that Nancy is certain that Zoe calculated the tip?

impossible **definitely**

Next

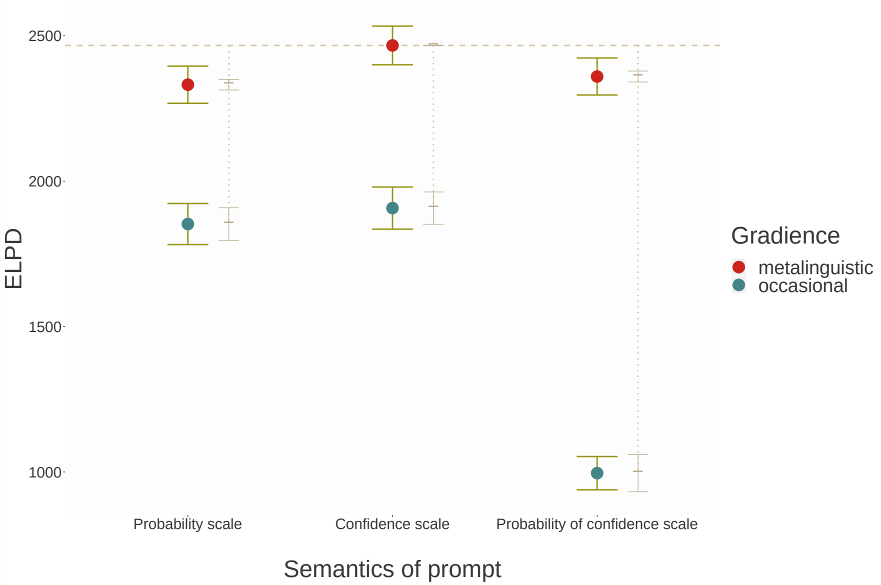
- Start with a prior distribution over *discourse states.*
  Update with ⟦*Zoe is a math major*⟧.
  Update with ⟦*Tim knows that Zoe calculated the tip*⟧.
- Push ⟦*How likely is it that N is certain that Z calculated the tip?*⟧ onto the QUD stack.
- Pop it off the stack; respond.

Semantics of '*how likely is it that X is certain that p*'

- Assumption: *certain* gives rise to a vague standard threshold, and thus *occasional* uncertainty.
  likely computes the probability of the vague inference.
- Perhaps, more appropriate to think of this standard as being *imprecise* rather than *vague*. . . .
  assuming *certain* is a maximum standard adjective (see, e.g., Kennedy (2007) and Kennedy and McNally (2005)).

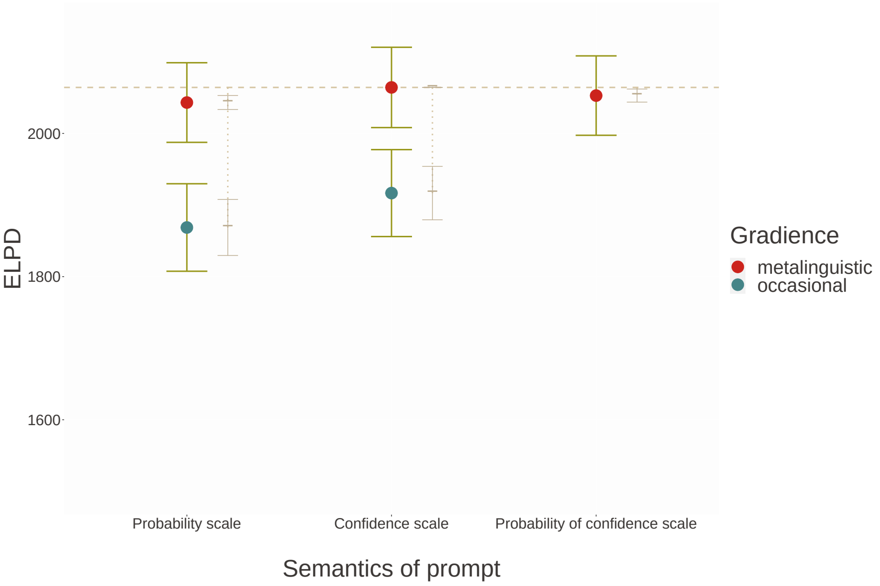
Results

The “how certain” task: comparing models



2500

The “how likely … certain” task: comparing models



Summing up

Outlook

- Building on Grove and White (2024), we continue to find that inferences from factive predicates exhibit gradience which is metalinguistic in nature.
- When incorporating semantic analyses into our probabilistic models, there is an advantage to going all the way!
- The compositional semantics of the prompt matters.
- Probabilistic dynamic semantics allows us to seamlessly incorporate it into our models of inference data.

- Degen, Judith and Judith Tonhauser (2021). **"Prior Beliefs Modulate Projection".** In: *Open Mind* 5, pp. 59–70. DOI: 10.1162/opmi_a_00042.
- Degen, Judith and Judith Tonhauser (2022). **"Are there factive predicates? An empirical investigation".** en. In: *Language* 98.3. Publisher: Linguistic Society of America, pp. 552–591. DOI: 10.1353/lan.0.0271.
- Djärv, Kajsa and Hezekiah Akiva Bacovcin (2017). **"Prosodic Effects on Factive Presupposition Projection".** en. In: *Semantics and Linguistic Theory* 27.0. Number: 0, pp. 116–133. DOI: 10.3765/salt.v27i0.4134.

References ii

Djärv, Kajsa, Jérémy Zehr, and Florian Schwarz (2018). **"Cognitive vs. emotive factives: An experimental differentiation".** en. In: *Proceedings of Sinn und Bedeutung*. Vol. 21. Number: 1, pp. 367–386.

Farkas, Donka F. and Kim B. Bruce (2010). **"On Reacting to Assertions and Polar Questions".** In: *Journal of Semantics* 27.1, pp. 81–118. DOI: `10.1093/jos/ffp010`.

Grove, Julian and Aaron Steven White (2024). ***Factivity, presupposition projection, and the role of discrete knowlege in gradient inference judgments.*** LingBuzz Published In: submitted.

- Kennedy, Christopher (2007). **"Vagueness and grammar: the semantics of relative and absolute gradable adjectives".** en. In: *Linguistics and Philosophy* 30.1, pp. 1–45. DOI: 10.1007/s10988-006-9008-0.
- Kennedy, Christopher and Louise McNally (2005). **"Scale Structure, Degree Modification, and the Semantics of Gradable Predicates".** In: *Language* 81.2. Publisher: Linguistic Society of America, pp. 345–381. DOI: 10.1353/lan.2005.0071.
- Klecha, Peter (2012). **"Positive and Conditional Semantics for Gradable Modals".** en. In: *Proceedings of Sinn und Bedeutung* 16.2. Number: 2, pp. 363–376.

- Roberts, Craige (2012). **"Information Structure: Towards an integrated formal theory of pragmatics".** en. In: *Semantics and Pragmatics* 5, 6:1–69. DOI: 10.3765/sp.5.6.
- White, Aaron Steven and Kyle Rawlins (2018). **"The role of veridicality and factivity in clause selection".** In: *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society*. Ed. by Sherry Hucklebridge and Max Nelson. Vol. 48. University of Iceland: GLSA (Graduate Linguistics Student Association), Department of Linguistics, University of Massachusetts, pp. 221–234.

Appendix A: probabilistic dynamic semantics

Ingredients

Typed λ-calculus

$$\frac{}{\Gamma, x : \alpha \vdash x : \alpha}\ \mathsf{Ax} \qquad \frac{\Gamma, x : \alpha \vdash t : \beta}{\Gamma \vdash \lambda x.t : \alpha \to \beta}\ {\to}\mathrm{I} \qquad \frac{\Gamma \vdash t : \alpha \to \beta \qquad \Gamma \vdash u : \alpha}{\Gamma \vdash t(u) : \beta}\ {\to}\mathrm{E}$$

$$\frac{}{\Gamma \vdash \diamond : \diamond}\ \diamond\mathrm{I} \qquad \frac{\Gamma \vdash t : \alpha \qquad \Gamma \vdash u : \beta}{\Gamma \vdash \langle t, u \rangle : \alpha \times \beta}\ \times\mathrm{I} \qquad \frac{\Gamma \vdash t : \alpha_1 \times \alpha_2}{\Gamma \vdash \pi_i(t) : \alpha_i}\ \times\mathrm{E}$$

Probabilistic programs

$$\frac{\Gamma \vdash t : \alpha}{\Gamma \vdash \boxed{t} : \mathsf{P}\alpha}\ \mathsf{Return} \qquad \frac{\Gamma \vdash t : \mathsf{P}\alpha \qquad \Gamma, x : \alpha \vdash u : \mathsf{P}\beta}{\Gamma \vdash \begin{pmatrix} x \sim t \\ u \end{pmatrix} : \mathsf{P}\beta}\ \mathsf{Bind}$$

Example: *tall*

(1) Jo is tall.
$\rightsquigarrow$ Jo's height exceeds some contextually salient threshold.

$$[\![\textit{tall}]\!] = \begin{pmatrix} d \sim \text{thresholdPrior} \\ \boxed{\lambda x.\text{height}(x) \geq d} \end{pmatrix} : \mathsf{P}(e \to t)$$

$$[\![\textit{Jo is tall}]\!] = \begin{pmatrix} d \sim \text{thresholdPrior} \\ \boxed{\text{height}(\text{j}) \geq d} \end{pmatrix} : \mathsf{P}t$$

Appendix B: models, more formally

Norming task (Degen and Tonhauser 2021)

Fact: Zoe is 5 years old.

How likely is it that Zoe calculated the tip?

impossible definitely

Continue

- Slider endpoints denote bounds of the scale for *likely*.

$\mu \sim \texttt{prior}$

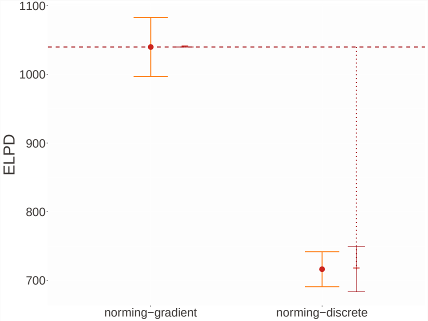
$\mu' \sim \text{update}_{\text{cg}}(\lambda w.[\![\textit{Zoe is 5 y.o.}]\!]^{w,\mu})(\mu)$

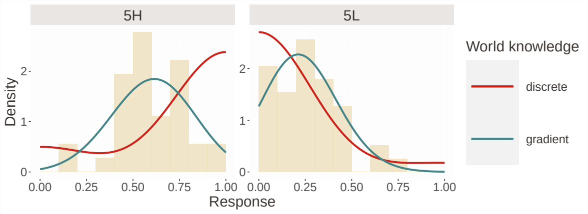
$x \sim \boxed{\max(\lambda d.\text{likely}(d)(\text{cg}(\mu'))(\lambda w.[\![\textit{Zoe calculated the tip}]\!]^{w,\mu'}))}$

$\boxed{\mathcal{N}(x, \sigma)\text{T}[0, 1]}$

Which does the truth of *Zoe calculated the tip* depend on: w, or μ'?

Norming task: comparing models





Experiment 1: the "how certain" task

Fact (which Nancy knows): Zoe is a math major..

Nancy asks: "*Does Tim know that Zoe calculated the tip?*"

How certain is Nancy that Zoe calculated the tip?

not at all certain **completely certain**

Next

$$\mu \sim \texttt{prior}$$

$$\mu' \sim \texttt{update}_{\text{cg}}(\lambda w.[\![\textit{Zoe is a math major}]\!]^{w,\mu})(\mu)$$

$$\mu'' \sim \texttt{update}_{\text{cg}}(\lambda w.[\![\textit{Tim knows Zoe calculated the tip}]\!]^{w,\mu'})(\mu')$$

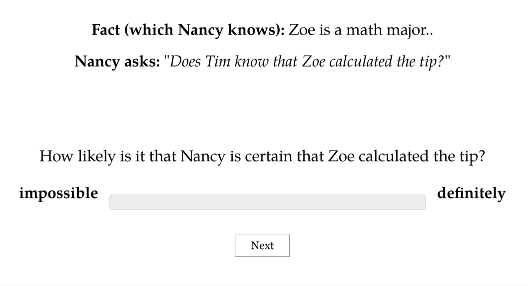
$$x \sim \boxed{\max(\lambda d.\text{certain}(d)(\text{cg}(\mu''))(\lambda w.[\![\textit{Zoe calculated the tip}]\!]^{w,\mu''}))}$$

$$\mathcal{N}(x, \sigma)\text{T}[0, 1]$$

Which does the truth of *Zoe calculated the tip* depend on:

w only? Or both w and μ''?

Experiment 2: the “how likely … certain” task



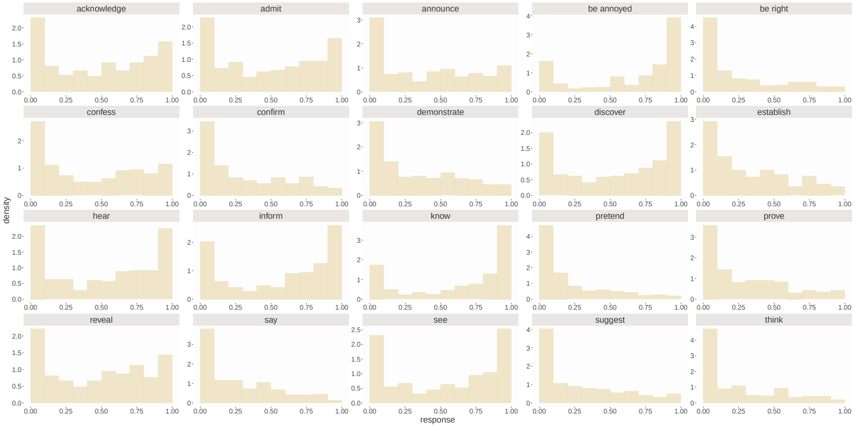
$$\mu \sim \texttt{prior}$$
$$\mu' \sim \texttt{update}_{\text{cg}}(\lambda w.[\![\textit{Zoe is a math major}]\!]^{w,\mu})(\mu)$$
$$\mu'' \sim \texttt{update}_{\text{cg}}(\lambda w.[\![\textit{Tim knows Zoe calculated the tip}]\!]^{w,\mu'})(\mu')$$
$$x \sim \boxed{\max(\lambda d.\text{likely}(d)(\text{cg}(\mu''))(\lambda w.[\![\textit{certain that Z calculated the tip}]\!]^{w,\mu''}))}$$
$$\mathcal{N}(x, \sigma)\texttt{T}[0, 1]$$

Which does the truth of *Zoe calculated the tip* depend on:

w only? Or both w and μ''?

Appendix C: data

Experiment 1: the “how certain” task



Experiment 2: the “how likely … certain” task