Proceedings of ELM 3: 176-187, 2025

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 comparisons couched within the framework of probabilistic dynamic semantics, we find that probabilistic models structured, in part, by the semantics of the prompt fit better to data collected using that prompt than models that ignore the semantics of the prompt.

Keywords. presupposition, factivity, dynamic semantics, probabilistic models

1. Introduction. When collecting inference judgments in formal experiments, it is common for trials to consist of the following pieces: (i) a target linguistic expression whose inferential affordances one is interested in measuring; (ii) a description of some context which the expression might be used in; (iii) a natural language prompt guiding participants on the relevant task; and (iv) a response instrument, such as an ordinal or slider scale. When analyzing the data from such experiments, one generally incorporates some representation of components (i), (ii), and (iv); but it is relatively rare to incorporate a representation of component (iii)—likely because this component is typically constant across all experimental items.[1]

We show that it can be important to incorporate an explicit representation of the semantics of natural language prompts used in inference judgment experiments into one's data analysis regime. To demonstrate this, we fix components (i), (ii), and (iv) of an experimental paradigm focused on measuring factive inferences—such as the inference from (1a) to (1b)—and manipulate component (iii)—the natural language prompt—in small but semantically potent ways.

(1) a. Jo {loves, doesn't love} that Mo left.
    b. Mo left.

We collect data using two such manipulated prompts and show, through statistical model comparison, that probabilistic models structured, in part, by the semantics of the prompt fit better to data collected using that prompt than models that ignore its semantics.

In Section 2, we describe how we incorporate natural language prompts into statistical models using the framework of probabilistic dynamic semantics (PDS; Grove & White 2024a,b). We then review, in Section 3, the experimental paradigm and prior models within PDS of data collected

*We thank the organizers of ELM 3, as well as four anonymous reviewers. Authors: Julian Grove, University of Rochester (julian.grove@gmail.com) & Aaron Steven White, University of Rochester (aaron.white@rochester.edu).

[1]One might conceive of the way that hypotheses are modeled in the natural language inference (NLI) literature (Cooper et al. 1996, Dagan, Glickman & Magnini 2006, MacCartney 2009, *et seq.*) as an exception; but most NLI models assume a single ground truth label aggregated from possibly multiple participants responses (cf. Gantt, Kane & White 2020), rather than modeling the distribution of reponses themselves, as we do here.

under that paradigm. In Section 4, we describe our two modifications to this paradigm and the two corresponding experiments. In Section 5, we describe our model implementations and report model comparisons before concluding in Section 6.

2. Probabilistic dynamic semantics. We couch our modeling within the recently developed probabilistic dynamic semantics (PDS) framework (Grove & White 2024b). One of the core ideas we implement within this framework is that an inference judgment task may be characterized as a kind of discourse: a target sentence and its context of interpretation may be analyzed as updates to a common ground; meanwhile, the prompt may be regarded as contributing a new question under discussion (QUD; Ginzburg 1996, Roberts 2012). The upshot is that once we fix these formal components, we need only make linking assumptions—i.e., assumptions about how the probability distribution over possible answers to the QUD (given a common ground, appropriately updated) should manifest as a particular distribution of responses, given some response instrument. Our approach is therefore analogous to standard analysis regimes, insofar as one is always making linking assumptions via one's choice of statistical model (see Jasbi, Waldon & Degen 2019), but it has the additional benefit of providing an explicit interface between the formal semantics of an experimental item and such linking assumptions. We provide certain crucial details below. For a full introduction to PDS, see Grove & White 2024b.

2.1. Distributions on discourse constructs. In PDS, an ongoing discourse is represented by a map from an input state to a probability distribution over output states. States are simply tuples of *metalinguistic parameters*; these include, e.g., the common ground (Stalnaker 1978) and the QUD, as well as other conversationally relevant features of discourse (e.g., possible antecedents for elliptical constructions). One can view the state as akin to the context state of Farkas & Bruce (2010); our state, however, is less constrained in principle, in the sense that it is compatible with arbitrary representations of the context.[2] Notating the types of these state tuples 'σ', 'σ'', etc., an ongoing discourse is of type $\sigma \to \mathtt{P}\sigma'$, where $\mathtt{P}\sigma'$ is the type of *probability distributions* over states of type σ'. (More generally, $\mathtt{P}\alpha$ is the type of probability distributions over values of type α.) In words, a discourse is a map from input states to probability distributions over output states.

To perform updates of various types to some ongoing discourse, we use two operators—*bind* and *return*—which allow probability distributions to be sampled (bind) and ordinary, non-probabilistic values to be returned as *degenerate* distributions over those values (return).[3] To illustrate, suppose we have some categorical distribution $\mathsf{mammal} : \mathtt{P}e$ on mammals (where e is the type of entities). We can represent a distribution on mammals' mothers as:

$$\begin{array}{l} x \sim \mathsf{mammal} \\ \boxed{\mathsf{mother}(x)} \end{array}$$

[2]Furthermore, as Grove & Bernardy (2023) show, typed frameworks for probabilistic semantics can be used to encode Rational Speechs Acts (RSA) models, as laid out in Frank & Goodman 2012, Goodman & Frank 2016. Frameworks like PDS diverge from common implementation assumptions in such settings, however, in the sense that they have strong commitments to computational purity that often do not hold in practical applications of RSA (for discussion, see Grove & Bernardy 2023, as well as Grove & White 2024b).

[3]Importantly, these operators conform to the monad laws. See Grove & Bernardy 2023, Grove & White 2024b for why this fact is useful.

Here, $\mathsf{mother} : e \to e$ maps an entity to its mother. Meanwhile, x is sampled from mammal using bind, and $\mathsf{mother}(x)$ is then *returned*, as notated by the orange box. The idea is that this distribution assigns to any given entity y the probability $\sum_{x \in \mathsf{mother}^{-1}(y)} p_{\mathsf{mammal}}(x)$ (where $\mathsf{mother}^{-1} : e \to e \to t$ is the inverse of the mother function and $p_{\mathsf{mammal}}(x)$ is the probability that mammal assigns to x).

In a similar fashion, given two discourses $D_1 : \sigma \to \mathtt{P}\sigma'$ and $D_2 : \sigma' \to \mathtt{P}\sigma''$, we can invoke bind to sequence D_1 with D_2 and obtain a new discourse $D_3 : \sigma \to \mathtt{P}\sigma''$:

$$D_3 = \lambda s. \begin{pmatrix} s' \sim D_1(s) \\ D_2(s') \end{pmatrix}$$

Somewhat metaphorically: given a starting state s, we sample an output state from $D_1(s)$ and apply D_2 to this output state. Moreover, generalizing the previous example, given an input state s, the probability $p_{D_3(s)}(s'')$ that $D_3(s)$ assigns to an output state s'' is computed as:

$$p_{D_3(s)}(s'') = \sum_{s'} p_{D_2(s')}(s'') * p_{D_1(s)}(s')$$

In words, sequencing two maps from input states to probability distributions over output states involves doing pointwise multiplication of the probabilities assigned to intermediate states by the first map and those assigned to an output state by the second map and then summing the result.

Finally, we represent the common ground as a probability distribution over indices representing information about possible worlds, along with certain linguistic parameters (which, for current purposes, we keep fairly open ended); we refer to the type of indices as 'ι', so that common grounds themselves are distributions of type $\mathtt{P}\iota$. Given some index $i : \iota$, we refer to the "world" of i as 'w_i' and to what we call the "context" of i as 'c_i'; contexts are what we use to encode useful linguistic information. The idea is that worlds w determine, e.g., how tall a particular individual is, while contexts c determine, e.g., the vague threshold of height past which one's height is considered tall. Thus while a world-sensitive function such as $\mathsf{height}(w_i) : e \to r$ intuitively says something about how tall an entity is at an index i, the adjective's threshold $\mathsf{d}_{\mathit{tall}}(c_i) : r$ only provides information about an aspect of the lexical semantics of *tall* at that index, viz., what is required of an entity to count as tall.

2.2. MAKING AN ASSERTION. Similar to discourses, the meanings of expressions are both probabilistic and dynamic. Accordingly, we represent them as functions of type $\sigma \to \mathtt{P}(\alpha \times \sigma')$: given an input state, the meaning of an expression produces a probability distribution over *pairs* of ordinary meanings of type α and possible output states. Furthermore, given a sentence whose probabilistic dynamic meaning ϕ is of type $\sigma \to \mathtt{P}((\iota \to t) \times \sigma')$, we can represent an *assertion* of that sentence as a discourse which updates the common ground. Specifically, we have a function $\mathtt{assert}$ with the following type:

$$\mathtt{assert} : (\sigma \to \mathtt{P}((\iota \to t) \times \sigma')) \to \sigma \to \mathtt{P}\sigma'$$

Given such a ϕ, $\mathtt{assert}(\phi)$ is a discourse of type $\sigma \to \mathtt{P}\sigma'$ representing an assertion of ϕ. For any input state $s : \sigma$, $\mathtt{assert}(\phi)(s)$ samples a proposition p together with an output state s' from ϕ and simply updates the common ground of s' with p; in particular, it gives back a new state s'' just like s', except that the common ground of s'' ($\mathsf{CG}(s'')$) updates the common ground of s' ($\mathsf{CG}(s')$) with p. Ultimately, assertions modify an ongoing discourse so that its probability distribution over output states involves common grounds in which the proposition returned by ϕ has been observed to hold true (see Grove & White 2024b for details).

2.3. ASKING A QUESTION. We follow a categorial tradition by analyzing questions as denoting—given an index—sets of true short answer meanings (Hausser & Zaefferer 1978, Hausser 1983, Xiang 2021; cf. Karttunen 1977, Groenendijk & Stokhof 1984). For simplicity (but not by necessity), we assume that all questions are degree questions: given index, they denote sets of degrees of type r (real numbers). Questions therefore have probabilistic dynamic meanings of type $\sigma \rightarrow \mathtt{P}((r \rightarrow \iota \rightarrow t) \times \sigma')$.

Asking a question is a matter of pushing a question meaning onto the QUD stack (Roberts 2012, Farkas & Bruce 2010). Reflecting this, there is a function `ask` with the following type:

$$\texttt{ask} : (\sigma \rightarrow \mathtt{P}((r \rightarrow \iota \rightarrow t) \times (\sigma'_1 \times \delta \times \sigma'_2))) \rightarrow \sigma \rightarrow \mathtt{P}(\sigma'_1 \times ((r \rightarrow \iota \rightarrow t) \times \delta) \times \sigma'_2)$$

Given a probabilistic dynamic question meaning κ and an input state s, $\texttt{ask}(\kappa)(s)$ samples a pair of a question meaning $q : r \rightarrow \iota \rightarrow t$ and a state $s' : \sigma'_1 \times \delta \times \sigma'_2$ from $\kappa(s)$ and then updates the QUD stack of s' (something of type δ) by pushing q onto it. Thus it returns a new output state of type $\sigma'_1 \times ((r \rightarrow \iota \rightarrow t) \times \delta) \times \sigma'_2$.

2.4. RESPONDING TO A QUESTION. PDS also models responses to questions; at any point in an ongoing discourse, one can respond to the QUD at the top of the current QUD stack based on one's prior knowledge. Concretely, a given responder has some background knowledge $bg : \mathtt{P}\sigma$ constituting a prior distribution over *starting* states. The responder uses this prior, in conjuction with the interim updates to the discourse, to derive a probability distribution over answers to the QUD. This answer distribution is gotten by retrieving the QUD of any given state s'—resulting in a distribution over QUDs—and then taking the maximum degree of which it is true at an index sampled from the common ground—resulting, finally, in a distribution over degrees.[4]

2.5. LINKING ASSUMPTIONS. In practice (e.g., in the setting of a formal experiment), an answer needs to be given using a particular testing instrument. We assume that a given testing instrument may be modeled by a family f of distributions representing the likelihood, which is then fixed by a collection Φ of nuisance parameters. Thus we may define a family of *response functions*, parametric in the testing instrument (i.e., likelihood function), each of which takes a distribution bg representing one's background knowledge, along with an ongoing discourse m:

$$\texttt{respond}^{f_\Phi : r \rightarrow \mathtt{P}\rho} : \mathtt{P}\sigma \rightarrow (\sigma \rightarrow \mathtt{P}\sigma') \rightarrow \mathtt{P}\rho$$

For a fixed likelihood function f_Φ mapping any given real number answer onto a distribution over possible responses of type ρ (for some ρ), the response function takes a distribution representing background knowledge and a discourse to produce a response distribution. It does this by composing the discourse with background knowledge, as above, and then obtaining the maximum degree answer to the current QUD, before applying the likelihood function f_Φ to this degree.

[4]See Grove & White 2024b for details.

The testing instrument employed in our experiments, for example, is a slider scale that records responses on the unit interval $[0, 1]$. A suitable likelihood would therefore be a truncated normal distribution: $f(x, \Phi) = \mathcal{N}(x, \sigma)\,\mathsf{T}[0, 1]$ (so that $f = \mathcal{N}$ and $\Phi = \sigma$). This likelihood—which Grove & White (2024a) also employ in their models of factivity—can be viewed as allowing some distribution of response errors, given the intended target response (i.e., the answer to the question).

These are the ingredients we need to model the effects of the fine-grained semantics of the target and question prompt, given a particular inference task. We provide further details in Section 5.

3. Factive inferences. To investigate how natural language prompts modulate the distribution of participants responses, we modify an experimental paradigm developed by Degen & Tonhauser (2021, 2022), which they use to experimentally investigate the projective inferences triggered by factive predicates—henceforth, *factive inferences*. We describe the paradigm (Section 3.1), then discuss previous modeling of their data within PDS (Section 3.2). The paradigm forms the basis for our experiments in Section 4; we build directly on our previous modeling in Section 5.

3.1. MEASURING FACTIVE INFERENCES. Degen & Tonhauser's (2021) main aim is to characterize the influence of world knowledge on factive inferences. To achieve this, they measure factive inferences in the presence of a background fact whose content they manipulate (their experiment 2b). For example, participants are given trials of the form in (2) and are asked to respond using a slider scale, with *no* on one end and *yes* on the other.

(2) a. **Fact (which Elizabeth knows):** Zoe is a math major.
 b. **Elizabeth asks:** "Does Tim know that Zoe calculated the tip?"
 c. Is Elizabeth certain that Zoe calculated the tip?

They focus on the set of twenty clause-embedding predicates listed in (3).[5]

(3) Twenty clause-embedding predicates (Degen & Tonhauser 2022, p. 559, ex. 13)
 a. canonically factive: *be annoyed*, *discover*, *know*, *reveal*, *see*
 b. non-factive
  (i) non-veridical non-factive: *pretend*, *say*, *suggest*, *think*
  (ii) veridical non-factive: *be right*, *demonstrate*
 c. optionally factive: *acknowledge*, *admit*, *announce*, *confess*, *confirm*, *establish*, *hear*, *inform*, *prove*

Each embedded clause in their experiment is paired with one of two facts: either a fact intended to make the clause likely to be true (as in the example above), or a fact intended to make the clause unlikely to be true. To validate the use of these background facts for this purpose, Degen & Tonhauser (2021) conduct a norming experiment, in which the prior certainties about the truth of the complement clauses featured in their projection experiment are assessed independently, given

[5] The grouping in (3) is due to Degen & Tonhauser 2022 and is based on the prior literature on factivity (Kiparsky & Kiparsky 1970, Karttunen 1971, *et seq.*).

the same background facts (their experiment 2a). Trials in this experiment ask participants to judge how likely the relevant clause is to be true, given one of the two background facts constructed for it. For example, participants are given trials of the form in (4) and are asked to respond using a slider scale, with *impossible* on one end and *definitely* on the other.

(4) a. **Fact:** Zoe is a math major.
    b. How likely is it that Zoe calculated the tip?

Degen & Tonhauser find that the by-item means for the forty pairs of complement clauses and background facts, as assessed in their norming experiment, are a good predictor of the inference ratings for items featuring the same complement clauses and facts which they obtain in their experiment investigating projective inferences.

3.2. MODELING FACTIVE INFERENCES. Degen & Tonhauser (2022) and others (White & Rawlins 2018) observe that measures of a predicate's factivity derived from the sort of judgment data Degen & Tonhauser (2021) collect display gradience when the data is aggregated.[6]

Using models developed in PDS, Grove & White (2024a) ask whether this apparent gradience arises due to *metalinguistic uncertainty*—uncertainty about whether a predicate is factive or not—or *occasional uncertainty*—uncertainty inherently associated with predicate meanings.[7] Under a metalinguistic uncertainty account, different predicates differ in the frequencies with which they trigger factive inferences across uses. Under an occasional uncertainty account, predicates would license inferences with varying degrees of certainty on particular uses, similar to the manner in which a vague predicate, such as *tall*, can license uncertain inferences about the heights of individuals of which it is predicated.

Grove & White fit four models to Degen & Tonhauser's data, varying whether uncertainty about either background world knowledge or factivity is encoded as metalinguistic or occasional. We extend their models here by adding an explicit model of the semantics of the natural language prompt. In Degen & Tonhauser's (2021) original paradigm, this prompt is the one in (5).

(5) Is PERSON certain that CLAUSE?

Following Grove & White's suggestion that no extant proposal posits that world knowledge should display metalinguistic uncertainty—a suggestion consistent with their modeling results—we focus specifically on two of their models: the *discrete-factivity* model (DF), which regards uncertainty about factivity as metalinguistic and uncertainty about world knowledge as occasional, and the *wholly-gradient* model (WG), which regards both kinds of uncertainty as occasional.

Grove & White find that DF performs the best in a model comparison pitting all four of their models against each other, as assessed by expected log pointwise predictive densities (ELPDs).

[6] Degen & Tonhauser (2022) argue that this gradience is evidence that there is no distinct classes of factive predicates. This argument is based on an apparent lack of clear clustering in the aggregate measures of different predicates' ratings, as derived from inference judgment data collected using this and similar paradigms (White & Rawlins 2018, Ross & Pavlick 2019). This lack of clear clustering, however, is likely a product of measurement noise: when such noise is appropriately modeled, distinct classes of factive predicates reveal themselves (Kane, Gantt & White 2022).

[7] See Grove & White 2024a for the full formal details of their models.

They argue that this finding lends support to a view of factivity whereon it is a fundamentally discrete phenomenon, and they discuss how both conventionalist and conversationalist accounts might approach this sort of discreteness.

4. Modifying the prompt. While Grove & White's results are promising, they are consistent with the possibility that the nature of Degen & Tonhauser's prompt biased experimental participants toward making discrete 'yes' or 'no' judgments, even while the contribution to inference judgments made by factive predicates may be gradient. Because the prompt is a polar question, and *yes* and *no* label the slider scale, participants may effectively treat their response as a binary forced choice by providing an answer near *yes* if they are sufficiently certain about the relevant inference, and an answer near *no* if they are not. If so, an *a priori* advantage is conferred on models regarding the contribution to inference of factive predicates as discrete and, thus, models which regard uncertainty about factive inferences as metalinguistic.

To assess the effect the prompt has on participants' responses, we conduct two experiments identical to Degen & Tonhauser's, but which vary the prompt. In both, participants are provided with a *degree* question, which is either about the speaker's degree of certainty (6a) or the degree of *likelihood* that the speaker is certain (6b).

(6) a. How certain is PERSON that CLAUSE?

b. How likely is it that PERSON is certain that CLAUSE?

The prompt in (6a) was paired with a slider labeled *not at all certain* on the left and *completely certain* on the right, while the prompt in (6b) was paired with a slider labeled *impossible* and *definitely* (following Degen & Tonhauser's norming experiment).

The idea behind using the degree questions in (6), which involve the modal adjectives *certain* and/or *likely*, is that insofar as factivity is fundamentally gradient in nature, such degree questions should encourage participants to contact that fundamentally gradient representation, whatever it may consist in. At the very least, they should not discourage participants from relying on such a gradient representation (as a polar question might), and furthermore, they should not encourage them to discretize it. In PDS, these degree questions can be seen as driving inferences based on the gradience encoded in the common ground. If factivity is fundamentally gradient in nature, factive predicates should contribute correspondingly gradient updates.

All aspects of the experimental materials and methods (besides the prompts) match Degen & Tonhauser's. We recruited 300 participants to give judgments for each prompt. Data from 15 participants was removed in the experiment employing (6a), and data from 7 participants was removed in the experiment employing (6b); for both, we followed Degen & Tonhauser's criteria. Both groups of participants were recruited through Amazon Mechanical Turk and paid $2.

5. Modeling the prompt.

5.1. ADDING A PROMPT MODEL. We fit both the DF and WG models of Grove & White 2024a, while manipulating the semantics of the question prompt for both.[8] Specifically, to model the

[8]We specifically use the variants of Grove & White's models that employ a truncated normal likelihood and that incorporate an anti-veridicality component. See their appendix for details.

prompt in (6a), we assume that the degree introduced by *certain* ranges over degrees of confidence rather than degrees of probability (see Klecha 2012, Goodman 2023), and thus that its scale is truncated relative to that of *likely*. We refer to this model as the *confidence scale* model.

(7) $$[\![\textit{certain}]\!] = \lambda s.\ f \sim \boxed{\lambda\phi.Pr\left(\begin{array}{l} i \sim \mathsf{CG}(s) \\ \boxed{\phi(i)} \end{array}\right)} \\ \boxed{\langle \lambda\phi, d, i. \frac{\max(0, f(\phi) - \theta_{certain}(s))}{1 - \theta_{certain}(s)} \geq d, s\rangle}$$

The type of the expression in (7) is $\sigma \to \mathtt{P}(((\iota \to t) \to r \to \iota \to t) \times \sigma)$: it maps a proposition onto a question meaning that returns T for threshold degrees less than or equal to a degree of *confidence* obtained from the probability that the proposition is true in the common ground of the current state; in particular, *A is certain that S* is true at a threshold d if the probability of *S* is greater than d added to a fixed cutoff point determined by the current state, where this probability is further scaled by the size of the interval determined by the cutoff point.

To model the prompt in (6b), we assign a semantics to *likely* according to which it introduces a degree corresponding to a probability (as opposed to a degree of confidence; see (8)), and where this degree is computed based on the corresponding semantics for *certain*. We refer to this model as the *probability of confidence scale* model.

(8) $$[\![\textit{likely}]\!] = \lambda s. \boxed{\left\langle \lambda\phi, d, i.Pr\left(\begin{array}{l} i' \sim \mathsf{CG}(s) \\ \boxed{\phi(i')} \end{array}\right) \geq d, s \right\rangle}$$

Note that to compute the meaning of *likely* [*certain S*], the degree threshold of *certain* needs to be saturated, in order to obtain a proposition of type $\iota \to t$ which can be fed to *likely*. To accomplish this, we effectively assume two things: (a) that this threshold is a metalinguistic parameter; and (b) that the probability that the certainty about the prejacent *S* is greater than the threshold is determined by a distribution centered at the actual probability of *S*, which may vary by both participant and prejacent. Thus we effectively attribute to each experimental participant uncertainty about the epistemic state of the propositional attitude holder denoted by the relevant embedded subject; this uncertainty, furthermore, is represented by a distribution centered at a value estimated about the participant's background world knowledge. In the end, we effectively attribute to the bare form of *certain* the semantics of an (e.g., maximum standard) absolute gradable adjective.

Both models contrast with Grove & White's original model, which encodes a meaning for the prompt (5) according to which it asks for a value on a probability scale. We refer to this model as the *probability scale* model.

5.2. Model fitting. We fit all models using Hamiltonian Monte Carlo sampling as implemented in STAN—specifically, `CmdStanR` (Gabry & Češnovar 2023). Four chains were sampled for each model to assess convergence, with at least 6,000 warmup samples and at least 6,000 samples kept per chain. All convergence diagnostics implemented in `CmdStanR` were conducted. In cases where a model did not converge for reasons that can be solved by drawing more samples, the number of samples was increased until convergence was reached. Even after substantial increases

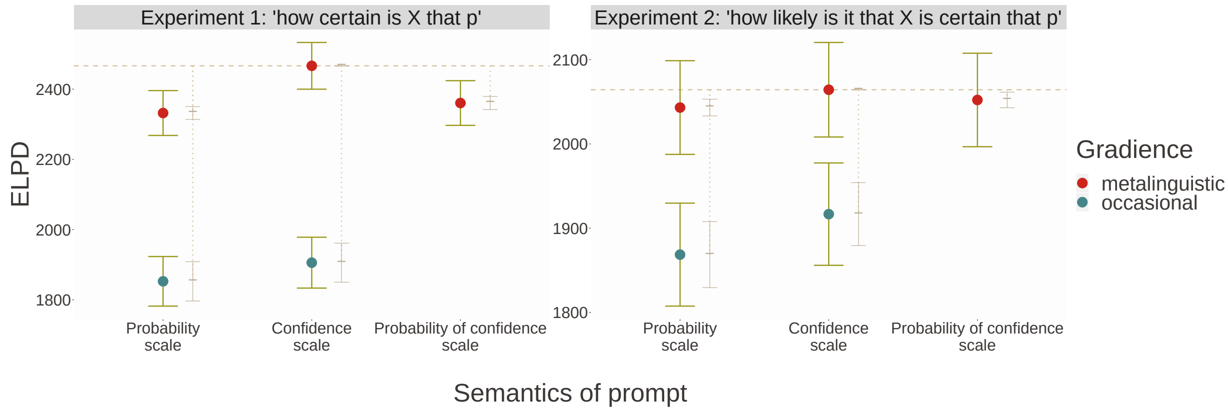

Figure 1: ELPDs for the models of the data collected using the prompts (6a) and (6b). We do not show the probability of confidence scale models encoding gradience as occasional in nature, since they do not converge. Error bars indicate standard errors of the ELPD, as well as of the pointwise differences from the best-performing model in each facet.

in the number of samples, we were unable to fit to convergence the probability of confidence scale models encoding gradience as occasional in nature. We do not show the results for this model here for this reason.[9]

5.3. Results. We follow Grove & White in reporting model comparisons using ELPD. Figure 1 shows the ELPD for the models of the data collected using the prompts (6a) and (6b). Error bars indicate standard errors of the ELPD, as well as of the pointwise differences from the best-performing model in each facet.

Recall that the DF models are those which regard factive inferences as giving rise to *metalinguistic* gradience, while the WG models are those regarding it as giving rise to *occasional* gradience. The y-axes of each plot provide the three variants of the semantics of the prompt discussed above.

Across the board, the DF models continue to perform the best on both datasets. Meanwhile, we find that confidence scale model performs the best on the dataset containing the prompt in (6a), as expected, while the confidence scale and probability of confidence scale models perform about equally on the dataset containing the prompt in (6b) and better than Grove & White's original prompt model (the probability scale model). This result might suggest that, while we may approximate the denotation of *likely* well, the denotation we assign to *certain* is not quite correct, even if it ends up providing the correct question meaning for the prompt in (6a). We take this result to highlight the benefits of developing models in the PDS framework, since it makes it straightforward to quantitatively evaluate alternative denotations, including those not considered in prior literature.

6. Conclusion. Our results suggest that, when analyzing data from an inference judgment task, it can be important to incorporate into one's data analysis regime an explicit representation of the

[9] The fits we obtained show extremely poor performance on both datasets, though we cannot read too much into this fact, given that these models did not converge.

semantics of the natural language prompt used to guide participants on the task. They additionally confirm (i) that the model comparisons obtained by Grove & White (2024a) do not reflect an *a priori* bias conferred on the discrete models by the experimental task, but rather these models' abilities to capture the distributions of degrees of certainty associated with the inferences generated by the predicates and complement clauses tested; and (ii) that while prior work has potentially provided a good approximation to the correct semantics for predicates like *certain* and *likely*, better approximations may potentially help in studying their semantic contributions to complex inferences involving both.

Future research in this line will aim to leverage PDS to explore the space of possible denotations for such predicates via explicit model comparison of the form we employ here. Beyond improving our understanding of the semantics of predicates like *likely*, we believe that the approach we have taken here can help us to better understand the fine-grained semantics of attitude predicates like *certain*, as well—a point supported by our success in modeling (6a). One clear opportunity for improving the denotation of predicates like *certain* is to investigate whether or not our current denotation takes the attitude holder's epistemic state into account in the right way: here we assume that the relevant contextual standard is fixed, while degrees of certainty vary according to a distribution determined by a given participant's background knowledge. Further investigating how an attitude holder's epistemic state should be incorporated into the semantics of *certain* is thus a potentially interesting future direction that is imminently feasible within PDS.

References

Cooper, Robin et al. 1996. *Using the Framework*. Technical Report LRE 62-051 D-16. The FraCaS Consortium.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. en. In Joaquin Quiñonero-Candela et al. (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11736790_9.

Degen, Judith & Judith Tonhauser. 2021. Prior Beliefs Modulate Projection. *Open Mind* 5. 59–70. https://doi.org/10.1162/opmi_a_00042. https://doi.org/10.1162/opmi_a_00042 (25 April, 2023).

Degen, Judith & Judith Tonhauser. 2022. Are there factive predicates? An empirical investigation. en. *Language* 98(3). Publisher: Linguistic Society of America, 552–591. https://doi.org/10.1353/lan.0.0271. https://muse.jhu.edu/article/864635 (28 November, 2022).

Farkas, Donka F. & Kim B. Bruce. 2010. On Reacting to Assertions and Polar Questions. *Journal of Semantics* 27(1). 81–118. https://doi.org/10.1093/jos/ffp010. https://doi.org/10.1093/jos/ffp010 (28 April, 2024).

Frank, Michael C. & Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336(6084). Publisher: American Association for the Advancement of Science, 998–998. https://doi.org/10.1126/science.1218633. https://www.science.org/doi/10.1126/science.1218633 (20 June, 2022).

Gabry, Jonah & Rok Češnovar. 2023. *CmdStanR*. Tech. rep. https://mc-stan.org/cmdstanr/index.html.

Gantt, William, Benjamin Kane & Aaron Steven White. 2020. *Natural Language Inference with Mixed Effects*. arXiv:2010.10501 [cs]. https://doi.org/10.48550/arXiv.2010.10501. http://arxiv.org/abs/2010.10501 (17 June, 2024).

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman & Dag Westerståhl (eds.), *Logic, Language, and Computation*, vol. 1, 221–237. Stanford: CSLI Publications.

Goodman, Jeremy. 2023. Degrees of confidence are not subjective probabilities. In *Proceedings of Sinn und Bedeutung 28*.

Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. en. *Trends in Cognitive Sciences* 20(11). 818–829. https://doi.org/10.1016/j.tics.2016.08.005. https://www.sciencedirect.com/science/article/pii/S136466131630122X (11 February, 2021).

Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies on the semantics of questions and the pragmatics of answers*. Amsterdam: University of Amsterdam dissertation. https://stokhof.org/wp-content/uploads/2020/09/groenendijk-stokhof_ssqpa.pdf.

Grove, Julian & Jean-Philippe Bernardy. 2023. Probabilistic Compositional Semantics, Purely. en. In Katsutoshi Yada et al. (eds.), *New Frontiers in Artificial Intelligence* (Lecture Notes in Computer Science), 242–256. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36190-6_17.

Grove, Julian & Aaron Steven White. 2024a. *Factivity, presupposition projection, and the role of discrete knowlege in gradient inference judgments*. LingBuzz Published In: submitted. https://ling.auf.net/lingbuzz/007450 (25 April, 2024).

Grove, Julian & Aaron Steven White. 2024b. *Probabilistic dynamic semantics*. LingBuzz Published In: https://ling.auf.net/lingbuzz/008478 (20 October, 2024).

Hausser, Roland & Dietmar Zaefferer. 1978. Questions and Answers in a Context-Dependent Montague Grammar. en. In F. Guenthner & S. J. Schmidt (eds.), *Formal Semantics and Pragmatics for Natural Languages*, 339–358. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9775-2_12. https://doi.org/10.1007/978-94-009-9775-2_12 (19 April, 2024).

Hausser, Roland R. 1983. The Syntax and Semantics of English Mood. en. In Ferenc Kiefer (ed.), *Questions and Answers*, 97–158. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-7016-8_6. https://doi.org/10.1007/978-94-009-7016-8_6 (19 April, 2024).

Jasbi, Masoud, Brandon Waldon & Judith Degen. 2019. Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate. English. *Frontiers in Psychology* 10. Publisher: Frontiers. https://doi.org/10.3389/fpsyg.2019.00189. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00189/full (3 July, 2024).

Kane, Benjamin, Will Gantt & Aaron Steven White. 2022. Intensional Gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising. en. *Semantics and Linguistic Theory*

31(0). 570–605. https://doi.org/10.3765/salt.v31i0.5137. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/31.029 (11 May, 2023).

Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4(1). Publisher: Routledge _eprint: https://doi.org/10.1080/08351817109370248, 55–69. https://doi.org/10.1080/08351817109370248.https://doi.org/10.1080/08351817109370248 (26 June, 2023).

Karttunen, Lauri. 1977. Syntax and Semantics of Questions. *Linguistics and Philosophy* 1(1). Publisher: Springer, 3–44. https://www.jstor.org/stable/25000027 (17 June, 2024).

Kiparsky, Paul & Carol Kiparsky. 1970. FACT. en. In *Progress in Linguistics*, 143–173. De Gruyter Mouton. https://doi.org/10.1515/9783111350219.143 (9 June, 2023).

Klecha, Peter. 2012. Positive and Conditional Semantics for Gradable Modals. en. *Proceedings of Sinn und Bedeutung* 16(2). Number: 2, 363–376. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/433 (14 December, 2023).

MacCartney, Bill. 2009. *Natural language inference*. en. Stanford: Stanford University dissertation. https://www-nlp.stanford.edu/~wcmac/papers/nli-diss.pdf.

Roberts, Craige. 2012. Information Structure: Towards an integrated formal theory of pragmatics. en. *Semantics and Pragmatics* 5. 6:1–69. https://doi.org/10.3765/sp.5.6. https://semprag.org/index.php/sp/article/view/sp.5.6 (26 July, 2023).

Ross, Alexis & Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2230–2240. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1228. https://aclanthology.org/D19-1228 (26 June, 2023).

Stalnaker, Robert. 1978. Assertion. In Peter Cole (ed.), *Pragmatics*, vol. 9, 315–332. New York: Academic Press.

White, Aaron Steven & Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In Sherry Hucklebridge & Max Nelson (eds.), *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society*, vol. 48, 221–234. University of Iceland: GLSA (Graduate Linguistics Student Association), Department of Linguistics, University of Massachusetts.

Xiang, Yimei. 2021. A hybrid categorial approach to question composition. en. *Linguistics and Philosophy* 44(3). 587–647. https://doi.org/10.1007/s10988-020-09294-8. https://doi.org/10.1007/s10988-020-09294-8 (26 September, 2024).