

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 out this investigation, we develop statistical models of presupposition projection that implement these two hypotheses, fit these models to existing inference judgment data aimed at measuring factive presuppositions (Degen and Tonhauser 2021), and compare the models’ fit to the data using standard statistical model comparison metrics. We find that models implementing the hypothesis that presupposition projection is fundamentally discrete fit the data better than models that implement the hypothesis that it is fundamentally gradient. To evaluate the robustness of this finding, we collect three additional datasets: a replication of the original dataset, as well as two datasets that modify the methodology of the original. Across each of these three datasets, we again find that models implementing the discreteness hypothesis fit the data better than models that implement the gradient hypothesis. Based on these results, we argue that classical semantic accounts of factive predicates can remain largely intact.

1 Introduction

Semantic theories aim to characterize the inferences that natural language expressions support and to account for at least a subset of the necessary inferences, given the meanings of the expressions. Whether or not a particular inference is necessary is commonly assessed via native speaker judgments. Judgment data, however, tends to be influenced by a number of non-semantic factors. These factors run the gamut: from high-level factors, such as speakers’ prior beliefs about the likelihood that an inference is true or ambiguities about the expressions involved, to low-level factors, such as the strategies speakers use to map their judgments to a data collection instrument (e.g., a slider representing likelihood or certainty) or their skill in producing an accurate target response using such instruments.

Testing a semantic theory against inference judgment data thus requires auxiliary assumptions about the link between (some representation of) these factors and the theoretical

*We thank Judith Degen and Judith Tonhauser for making their experimental data and materials available. Thanks to Lelia Glass and Florian Schwarz for useful comments on an earlier draft of this paper, and to Will Gantt for his tips on using Stan. Also thanks to Chris Kennedy and Ming Xiang and the participants of their Spring 2022 semantics seminar at the University of Chicago for feedback.

constructs of interest. Such *linking assumptions* are often left implicit in classical methodologies employing informal experiments. In recent years, however, the need to formulate explicit linking assumptions has become pressing in light of theoretical developments within semantics that are motivated by finer-grained aspects of the distribution of inference judgments than can be observed informally.

One area where such motivations have become particularly important is the domain of presupposition projection (in general) and factivity (in particular). A predicate is said to be *factive* if it is implicated in triggering *veridicality inferences*—i.e., inferences that the content of its embedded clause is true—regardless of whether or not entailment canceling operators take scope over that predicate (Kiparsky and Kiparsky 1970). For example, *love* is often taken to be factive, since sentences such as those in (1) give rise to the inference in (2).

- (1) a. Jo loves that Mo left.
 b. Jo doesn't love that Mo left.
 c. Does Jo love that Mo left?
 d. Jo might love that Mo left.
 e. If Jo loves that Mo left, she'll also love that Bo left.

- (2) Mo left.

Diagnosing factivity has long been known to be challenging due to the influence a predicate's context of use exerts on the relevant veridicality inferences (Karttunen 1971 *et seq*). Thus, to better understand the factors that drive factive inferences, it has become more common for researchers to collect judgments from native speakers in formal experiments, often in large quantities, in order to evaluate hypotheses about the semantic properties of factive predicates, as well as about how these semantic properties relate to the distributional properties of judgment data (Tonhauser 2016; Djärv and Bacovcin 2017; Djärv, Zehr, and Schwarz 2018; White and Rawlins 2018b; White, Rudinger, et al. 2018; White 2021; Degen and Tonhauser 2021; Degen and Tonhauser 2022; Jeong 2021; Kane, Gantt, and White 2022).

Of particular importance in the experimental literature on factivity has been the observation that, in tasks aimed at measuring a predicate's factivity, aggregate measures derived from inference judgment tasks show much more gradience than one might initially expect under a classical view of factivity as a discrete property (White and Rawlins 2018b). Some authors have gone so far as to claim that such gradience casts doubt on the very notion that there are discrete lexical properties driving factive inferences at all (Degen and Tonhauser 2022). Such doubt is consistent with the view that presupposition projection is fundamentally gradient in general (Tonhauser, Beaver, and Degen 2018). This *fundamental gradience hypothesis* contrasts with a *fundamental discreteness hypothesis*, which instead aims to retain the classical view of factivity as a discrete property (or collection of properties) by attributing a significant portion of the observed gradience to the sorts of non-semantic factors discussed above. We discuss these hypotheses in more detail in Section 2.

Our central aim in this paper is to quantitatively evaluate these two hypotheses by developing a framework that allows us to explicitly formulate the link between their respective construals of factivity and the way humans produce judgments that depend on these construals. The core theoretical contribution we make in developing this framework, which builds on one proposed by Grove and Bernardy (2023), is to provide a way to transparently relate the sorts of formal compositional analyses of expressions’ meanings that are common in the formal semantics literature to probabilistic models characterizing distributions over inference judgments.

We formally define the framework in Section 3 before using it to carry out an apples-to-apples comparison of the two hypotheses in Sections 4 and 5. This apples-to-apples comparison is possible because our framework allows us not only to precisely target where these hypotheses make different predictions about the distribution of inference judgments across participants, but furthermore does so by allowing us to use standard statistical model comparison metrics which balance out a model’s fit to some inference judgment data against the model’s complexity. Using such metrics, we find that models that implement the fundamental discreteness hypothesis unambiguously outperform models that implement the fundamental gradience hypothesis across both an existing dataset aimed at measuring factivity (Degen and Tonhauser 2021) and three new datasets: a replication of the existing dataset, along with two novel datasets. Based on these results, we argue in Section 6 that a classical semantic account of factive predicates can remain largely intact. We also discuss how the framework we develop in this paper might be understood as providing a common view of classical theories of factivity and theories that attempt to reduce it to an entirely pragmatic process (Simons 2007; Simons, Tonhauser, et al. 2010; Simons, Beaver, et al. 2017).

2 Gradient inference patterns among factive predicates

The advent of large-scale inference judgment datasets—such as MegaVeridicality (White and Rawlins 2018b; White, Rudinger, et al. 2018), VerbVeridicality (Ross and Pavlick 2019), and CommitmentBank (De Marneffe, Simons, and Tonhauser 2019)—has enabled fine-grained analyses of inference judgment patterns across the entire clause-embedding lexicon. Across such datasets, there is substantial gradience observed in the aggregate judgments of multiple speakers. This aspect of the distribution of inference judgments has garnered sustained focus.

In the domain of factivity, such gradience is noted by White and Rawlins (2018b), who look at aggregate measures of different predicates’ degree of factivity, using data from the MegaVeridicality dataset (see Figure 1). They observe that “there are not necessarily clear dividing lines between... classes [expected in a standard classification of clause-embedding predicates]... suggesting that speakers’ inferences about veridicality are generally quite gradient and likely influenced by the fine-grained semantics of particular verbs” (*ibid*, p. 228).¹

In later work building on White and Rawlins’s, Degen and Tonhauser (2022) investigate the nature of this gradience in six experiments, arguing that its persistence across experiments

¹This gradience is not White and Rawlins’s main focus, since they are interested in the relationship between inference and predicate distribution, rather than the semantic classification of predicates. They thus make no particular claims about its importance.

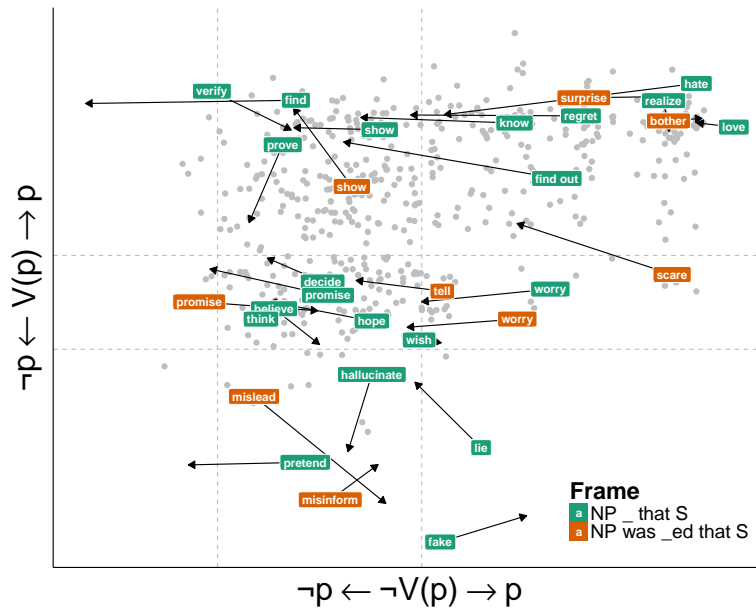


Figure 1: Aggregate measure of factivity derived from MegaVeridicality from White and Rawlins 2018b (their Ex. 14). The y -axis corresponds to an aggregate measure of responses to prompts of the form *Someone (was) __ed that a particular thing happened. Did that thing happen?* with possible responses *yes, maybe or maybe not, and no*. The x -axis corresponds to the same measure for prompts of the form *Someone {didn't, wasn't} __ that a particular thing happened. Did that thing happen?* with the same possible responses. Each grey point is a predicate, with a subset of predicates labeled. Label color corresponds to the syntactic context a predicate was measured in. Arrows and dotted lines can be ignored for current purposes.

militates against the hypothesis that there is a coherent class of factive predicates.

Our own modeling work uses data collected under the same experimental paradigm that Degen and Tonhauser employ, and so we describe their data and arguments in detail in Section 2.1. In Section 2.2, we turn to the broad question of which factors are responsible for the gradience observed in inference judgment tasks; we discuss evidence that, when one appropriately accounts for these factors, a small number of clear, inferentially defined classes of predicates are brought into relief (Kane, Gantt, and White 2022), thus casting doubt on Degen and Tonhauser's argument that there is no coherent class of factive predicates. Nonetheless, as we discuss in Section 2.3, there is apparent gradience internal to each of these classes, as well as among them, which may be compatible with the program, laid out by Tonhauser, Beaver, and Degen (2018), of viewing all presupposition projection as fundamentally gradient in nature. It is this latter hypothesis that we address in this paper.

2.1 Measuring veridicality and factivity

In each of their experiments, Degen and Tonhauser (2022) focus on the set of twenty clause-embedding predicates listed in (3), which they group into classes based on prior work (Kiparsky and Kiparsky 1970; Karttunen 1971; Hooper and Thompson 1973; Givón 1973; Hooper 1975; Abusch 2002; Abusch 2010; Abrusán 2011; Abrusán 2016; Anand and Hacquard 2014, i.a.).

- (3) Twenty clause-embedding predicates (Degen and 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*

In their discussion of the relationship between the traditional classification of these predicates and the experimental data involving projective inferences which they go on to collect, they say, “... we expect to see a categorical difference in projection between canonically factive predicates on the one hand, and optionally factive and nonfactive predicates on the other” (p. 569). To assess projection, Degen and Tonhauser provide participants with a scenario in which someone asks a polar question whose main verb is one of the factive predicates of interest, e.g., (4).

- (4) **Helen asks:** Did Amanda discover that Danny ate the last cupcake?

They then ask participants to provide a rating on a continuous scale from *no* to *yes* in answer to a prompt of the form in (5), in order to assess the extent to which participants believe that the embedded clause is presupposed.

- (5) Is Helen certain that Danny ate the last cupcake?

In another experiment, Degen and Tonhauser give participants a variant of this task in which their answer is provided as a binary forced choice between *no* and *yes*.

Degen and Tonhauser also claim that categories of predicates ought to emerge when analyzing judgments of veridicality: “we expect the [contents of the complements of] canonically factive and veridical nonfactive predicates to be entailed” (p. 569). They assess veridicality inferences using two methods. First, they provide participants with a scenario in which a sentence containing one of the predicates of interest is assumed to be true, as in (6).

- (6) **What is true:** Edward proved that Grace visited her sister.

They then prompt participants using a question of the form in (7).

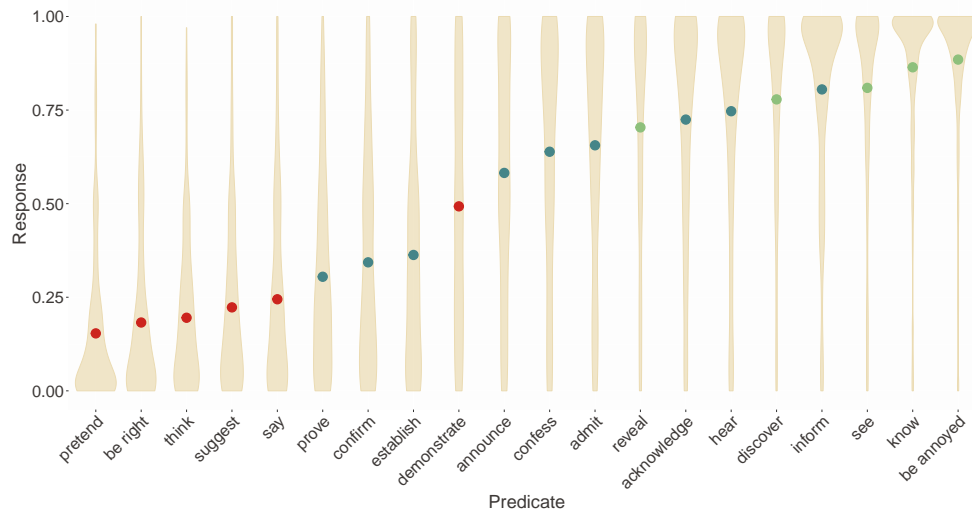


Figure 2: Verb means from Degen and Tonhauser’s (2022) experiment 1a. “Non-factive” verbs are in red, “optionally factive” verbs are in teal, and “canonically factive” verbs are in green. Violin plots indicate the probability density of responses. (See Degen and Tonhauser 2022, Figure 2, p. 562.)

(7) Does it follow that Grace visited her sister?

Depending on the experiment, either participants answer on a sliding scale from *no* to *yes*, or they are asked to make a binary forced choice between *no* and *yes*.

Second, Degen and Tonhauser provide participants with a scenario in which someone makes an utterance which should be contradictory if the relevant complement clause is entailed, as in (8).

(8) **Margaret:** “Edward heard that Mary is pregnant, but she isn’t.”

Participants are then prompted to answer a question of the form in (9) either on a sliding scale or by making a binary forced choice, depending on the experiment.

(9) “Is Margaret’s utterance contradictory?”

Consistent with White and Rawlins’s original observation, Degen and Tonhauser observe that the patterns of inference across predicates in the six experiments are gradient in nature for both projection and veridicality. The degree to which predicates display projective inferences appears to evolve continuously from the least projective predicate (*pretend*) to the most projective (*be annoyed*) when predicates are compared in terms of their mean ratings (see Figure 2). Such gradience is manifest in both of the experiments assessing projection—the one which collects sliding scale judgments and the one which collects binary judgments. A similar pattern emerges in the experiments assessing veridicality inferences. Crucially, no

predicate patterns consistently across all four of the experiments assessing veridicality with the control items that were constructed to generate entailments to the relevant clause.

2.2 Gradience in inference datasets

Degen and Tonhauser’s results are consistent, not only with White and Rawlins’s original observation, but with findings from adjacent domains. An and White (2020) observe similar gradience in neg-raising inferences captured in their MegaNegRaising dataset; and Kane, Gantt, and White (2022) note an analogous pattern among belief and desire inferences captured in their MegaIntensionality dataset.

Kane, Gantt, and White note that, in the face of such gradience, it is reasonable to entertain two kinds of hypotheses. One possibility is that “apparent gradience indicates that no formally represented lexical property controls whether a particular inference is triggered” (p. 572). Another is that “apparent gradience [may be] partly or wholly a product of the methods often used to collect inference judgments, and that there are discrete, formally represented lexical properties that are [nevertheless] active in triggering... inferences” (p. 572). To pursue this question, they ask whether clear *patterns* of inference emerge across the inference judgment datasets discussed above—MegaVeridicality, MegaNegRaising, and MegaIntensionality—by clustering predicates into classes according to the responses from those datasets so as to optimize their ability to predict predicates’ syntactic distributions, as measured in the MegaAcceptability dataset (White and Rawlins 2016).² They uncover fifteen classes of predicates that correspond extremely closely to those that one would expect from prior work on clause-embedding predicates. As can be seen in Figure 3, these classes include a variety of factive subclasses that differ principally in the pattern of belief and desire inferences they are associated with. As one might expect from prior literature, the true factive subclasses tend to be emotive, including, for example, *love* and *hate*.

Kane, Gantt, and White’s findings establish that there *is* a coherent class of factive predicates (which are, in turn, subclassed by the belief and desire inferences they give rise to). But they also find that there are a variety of classes associated with weaker veridicality inferences than one might expect from a truly factive class. These classes include non-emotive predicates, like *know* and *realize*. Thus while it is not correct to say that there is no class of factive predicates, one must still explain the source of the apparent gradience associated with certain classes, such as the non-emotive ones, as Degen and Tonhauser point out (Section 4.1, Objection 3). Class-level gradience of this kind is unlikely to be—as Kane, Gantt, and White put it—“partly or wholly a product of the methods often used to collect inference judgments”, since their analysis expressly accounts for the relevant task effects.

One way to account for this gradience is to adapt Tonhauser, Beaver, and Degen’s hypothesis that projection is fundamentally gradient to classes; for example, by admitting predicate classes that may be associated with particular amounts of gradience in the degree to which a predicate’s complement projects. Alternatively, one might take seriously a hypothesis which

²The idea behind optimizing the predictability of predicates’ syntactic distribution is that, insofar as the classes to which a predicate belongs are predictive of its syntactic distribution, there is preliminary evidence that that class is associated with some distributionally active lexical representation.

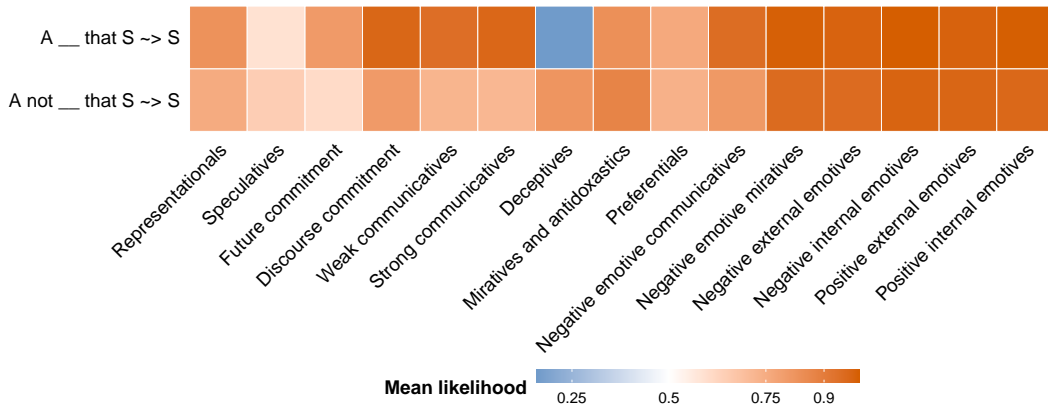


Figure 3: Veridicality inferences associated with each class found in Kane, Gantt, and White 2022. Kane, Gantt, and White provide labels for each class based on the predicates that occur in that class as well as the belief and desire inferences associated with that class. The top row and bottom row correspond to the y - and x -axes of Figure 1, respectively. The classes associated with dark orange cells in both rows are taken to be factive subclasses.

Degen and Tonhauser discuss—namely, that “the observed gradience in projection [is] compatible with a binary factivity category in combination with two assumptions: first, that predicates may be ambiguous between a factive lexical entry... and a nonfactive lexical entry... and, second, that interpreters may be uncertain about which lexical entry a speaker intended in their utterance” (p. 583). Our task in this paper is to formalize these two possibilities so that we may quantitatively compare them.

2.3 The role of world knowledge in gradient inference judgments

Our comparison will rely crucially on a paradigm used by Degen and Tonhauser (2021), who aim to characterize the influence of world knowledge on projection inferences, focusing on the same twenty clause-embedding predicates in (3). We employ this paradigm because it allows us to explicitly model the influence of non-semantic factors—specifically, speakers’ prior beliefs about the likelihood of an inference being true. Similar to the experiment reported above, in which Degen and Tonhauser (2022) measure presupposition projection out of the complement of a predicate placed inside of a polar question, Degen and Tonhauser (2021) measure projective inferences in the presence a background fact whose content they manipulate (their experiment 2b). To illustrate, the following experimental trial features the predicate *pretend*:

Fact (which Elizabeth knows): Zoe is a math major.
Elizabeth asks: "Did Tim pretend that Zoe calculated the tip?"

Is Elizabeth certain that Zoe calculated the tip?

no yes

Next

The same twenty complement clauses as from Degen and Tonhauser 2022 are also featured in this experiment, but now each clause 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. Each participant in this experiment sees twenty items (along with six control items). On each experimental trial, a predicate is placed in the context of one of the twenty clauses, along with one of the two background facts constructed for that clause. The results Degen and Tonhauser (2021) obtain in this setting mirror those of Degen and Tonhauser (2022). In particular, the mean projection ratings for the twenty predicates show a similar gradient pattern (Spearman’s $r = 0.98^{***}$).³

In addition to the assessment of projective inferences given background facts, Degen and 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 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, the following trial features the same clause as in the example given above, but with the alternate low-probability fact:

Fact: Zoe is 5 years old.

How likely is it that Zoe calculated the tip?

impossible definitely

Continue

Degen and 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 linear predictor of the inference ratings for items featuring the same complement clauses and facts

³Following standard convention, we use three asterisks to indicate that $p < 0.001$ when reporting correlation coefficients.

which they obtain in their experiment investigating projection inferences.⁴ Thus at least one source of variation among the projective inferences associated with clause-embedding predicates is the *context* in which these predicates are placed; in particular, the prior certainties that people associate with these contexts. This, of course, cannot be the whole story, as Degen and Tonhauser observe: the mean projection ratings for predicates display substantial gradience even after collapsing across the contexts in which they occur (following their 2022 experiment 1a: see Figure 2). So, what explains the remaining variation?

2.4 Two accounts of gradience

We consider two hypotheses about the source of variation in projective inference judgments among clause-embedding predicates:

- (10) a. *The Fundamental Discreteness Hypothesis.* Factivity is a discrete semantic property of at least some token occurrences of expressions containing at least some clause-embedding predicates. A given use of an expression containing a particular predicate either triggers a projective inference, or it does not trigger a projective inference.
- b. *The Fundamental Gradience Hypothesis.* There is no grammatical property distinguishing factive from non-factive occurrences of clause-embedding predicates. Rather, the gradient distinctions among predicates (and classes thereof) reflect the different gradient contributions specific predicates make to the inferences about the truth of their complement clauses.

According to the fundamental discreteness hypothesis, the gradience among the predicates discussed above is driven by metalinguistic knowledge. Individual occasions on which a predicate is used may be associated with uncertainty about whether or not the expression containing the predicate triggers projection, but that uncertainty is about which of the alternative interpretations of the expression should be selected. Alternative interpretations may be available because the predicate has multiple senses—at least one that is implicated in triggering projection and at least one that is not; or because the predicate may occur in multiple structures—at least one that is implicated in triggering projection and at least one that is not.⁵ The gradience associated with particular classes of predicates that Kane, Gantt, and White observe might then indicate a sort of regular polysemy among predicates within a class, or it might indicate that predicates in the class bias the resolution of syntactic ambiguity in similar ways.⁶

⁴They find a similar effect when the type of prior fact, whether “low” or “high”, is coded as a categorical variable.

⁵The second option is possible in light of a substantial amount of cross-linguistic evidence that functional items surrounding a predicate—e.g., nominal morphology attached to verbs or their clausal complements—can modulate veridicality inferences (see Varlokosta 1994; Giannakidou 1998; Giannakidou 1999; Giannakidou 2009; Roussou 2010; Farudi 2007; Abrusán 2011; Kastner 2015; Ozyildiz 2017; see White 2019 for discussion).

⁶Alternatively, one could posit that, rather than indeterminacy in the interpretation of an expression, there is uncertainty over possible questions under discussion (QUDs) against which the expression is interpreted. Such an alternative is compatible with, e.g. the proposal in Simons, Beaver, et al. 2017. While the QUD may be the

According to the fundamental gradience hypothesis, the variation in projective inferences among clause-embedding predicates is gradient because the inferences that the predicates trigger are themselves gradient (Tonhauser, Beaver, and Degen 2018). In this respect, such inferences may be on a par with those contributed by prior world knowledge: the use of a given predicate boosts the likelihood that its complement clause is true, but this boost is not conditioned by a discrete, formal aspect of the predicate’s semantic representation that produces a presupposition or an entailment. Crucially, there is no selection among alternative interpretations on particular occasions of use, under this hypothesis. One way to think about it is that it analogizes clause-embedding predicates, like *know*, to vague predicates, like *tall*.

Importantly, both of these hypotheses concern what comprehenders do when they draw inferences from particular uses of a predicate; that is, does their behavior look more like ambiguity resolution or more like, e.g., reasoning about vagueness? We stress that this question is independent of the question of whether there are factive predicates (or subclasses thereof), which we take to be resolved by Kane, Gantt, and White’s findings. Thus, our findings do not presuppose the correctness of Kane, Gantt, and White’s argument, though our interpretation of our own findings in Section 6 crucially makes reference to theirs.

3 Probabilistic semantics

To state a theory of gradience precisely, it is useful to have a general method for integrating probabilistic reasoning into a compositional semantics. Here, we rely on the framework provided by Grove and Bernardy (2023), which supplies an interface for performing Bayesian reasoning in the simply typed λ -calculus (with products) using *monads*.⁷ The main upshot of this framework is that it allows one to transparently relate the sorts of compositional analyses of expressions’ meanings common in the formal semantics literature to probabilistic models characterizing distributions over inference judgments.

As we will show, the framework allows for a precise specification of the two hypotheses laid out above, while keeping fixed both the formal analysis of the expressions of interest and the way in which probability distributions over inference judgments are mapped onto a particular data collection instrument. These aspects of the framework are important because, as we show in Sections 4 and 5, they allow us to conduct an apples-to-apples comparison of the two hypotheses that not only precisely targets where they make different predictions about the distribution of inference judgments across participants, but furthermore does so using standard statistical model comparison metrics which balance out a model’s fit to inference judgment data against the model’s complexity.

We begin in Section 3.1 with introductory background on Grove and Bernardy’s framework before turning in Section 3.2 to our extension of it, which allows us to finely delineate

source of ambiguity, it is not clear how to reconcile this approach with the observation that different (classes of) predicates are associated with particular levels of gradience without saying that lexical semantic knowledge conditions QUD choice. This move would violate the spirit of such *conversationalists* proposals, which generally attempt to do away with heavy conditioning on lexical information (see White 2019 for discussion).

⁷See Giorgolo and Asudeh 2014; Asudeh and Giorgolo 2020 and Bernardy, Blanck, Chatzikyriakidis, and Lappin 2018 for related monadic approaches. These approaches have slightly different aims from Grove and Bernardy’s, which are reflected in the distinct interfaces they supply.

uncertainty that is core to the semantic value of an expression—giving rise to phenomena such as vagueness—from uncertainty about which interpretation should be associated with a particular string—giving rise to metalinguistic uncertainty (see Bergen, Levy, and Goodman 2016; Potts et al. 2016; Monroe 2018 for a collection of related approaches). We illustrate the distinction between these two forms of uncertainty by first walking through an analysis of gradable adjectives in the Grove and Bernardy setting, since it allows us to highlight how their framework approaches vague predicates. We then give a minimalistic analysis of factivity in Section 3.3.

3.1 Denotations as probabilistic programs

The thrust of Grove and Bernardy’s framework is to provide an approach to *probabilistic semantics* that assimilates the probabilistic component of such a semantics to other notions of *effect* that have been studied in the formal semantics literature using monads—e.g., Shan’s (2002) first introduction of monads into this literature, with illustrations from focus, question semantics, anaphora, and quantification; Unger’s (2012) and Charlow’s (2014) approaches to anaphora using the State monad (and, in the latter case, the State transformer of Liang, Hudak, and Jones 1995); and various other phenomena, including conventional implicature (Giorgolo and Asudeh 2012), intensionality (Charlow 2020; Elliott 2022), and presupposition (Grove 2022).

To take an example familiar from probabilistic semantics settings, consider the meaning of the gradable adjective *tall*. Modeling only its role as a descriptor of individuals, one might regard *tall* as a predicate of type $e \rightarrow t$. To capture the contribution of *tall* to the entailments of expressions that contain it, one might then model its denotation as contributing the entailment that the height of the predicated individual x is greater than some contextually determined threshold d . Doing this, however, might lead to a semantic representation like the following, which involves an unbound degree variable d :

$$\lambda x.\text{height}(x) \geq d$$

There are different ways of remedying this situation. One approach assumes that the degree variable is existentially quantified—e.g., in virtue of the presence of an unpronounced morpheme which binds it—and that its value is constrained by some property made available by the context (see, e.g., Kennedy and McNally 2005). Another—and the one we build on here—leaves the variable unbound and relies on the context to directly fix its value (see, e.g., Barker 2002; Kennedy 2007).⁸ Among approaches that implement the second possibility, many rely on probabilistic knowledge to constrain how the value is fixed (Lassiter 2011; Goodman and Lassiter 2015; Lassiter and Goodman 2017; Bernardy, Blanck, Chatzikyriakidis, and Lappin 2018; Bernardy, Blanck, Chatzikyriakidis, Lappin, and Maskharashvili 2019a; Bernardy, Blanck, Chatzikyriakidis, Lappin, and Maskharashvili 2019b; Bernardy, Blanck, Chatzikyriakidis, and Maskharashvili 2022, i.a.).

Grove and Bernardy’s framework is one such probabilistic implementation, which uses a monad (P) to constrain the interpretation of the degree variable without tampering with the

⁸Our use of the term ‘variable’ here is a bit metaphorical: we mean to include any approach that values the standard of the relevant gradable adjective through contextual means.

underlying compositional semantics. \mathbb{P} maps types, such as e , t , $e \rightarrow t$, $e \times t$, etc., to types $\mathbb{P}e$, $\mathbb{P}t$, $\mathbb{P}(e \rightarrow t)$, $\mathbb{P}(e \times t)$, etc., which are inhabited by probabilistic programs. Because it is a monad, \mathbb{P} comes with two monadic operators: $(\sim) : \mathbb{P}\alpha \rightarrow (\alpha \rightarrow \mathbb{P}\beta) \rightarrow \mathbb{P}\beta$ (*bind*) and $(\boxed{\cdot}) : \alpha \rightarrow \mathbb{P}\alpha$ (*return*), which we describe in turn.

3.1.1 The ‘bind’ operator

The bind operator can be used to characterize the interpretation of contextually regulated parameters, like d above, by sequencing one probabilistic program with another that depends on a variable. This sequencing—notated $m \sim \lambda x.k(x)$ —can be understood as sampling a random value $x : \alpha$ from a probabilistic program $m : \mathbb{P}\alpha$, and then using that value to construct the new probabilistic program $k(x) : \mathbb{P}\beta$ (which is now parameterized by x). Following standard convention, $m \sim \lambda x.k(x)$ can be written in the following “imperative style”:

$$\begin{array}{l} x \sim m \\ k(x) \end{array}$$

We use this notation throughout the remainder of the paper. It is important to note that these two lines together describe the probabilistic program $m \sim \lambda x.k(x)$ and that similar multi-line descriptions below will also describe a single complex probabilistic program.

3.1.2 The ‘return’ operator

The return operator allows ordinary logical meanings to be *lifted* to probabilistic programs associated with a *trivial effect*.

$$\boxed{\cdot} : \alpha \rightarrow \mathbb{P}\alpha$$

The effect associated with the resulting program is trivial in the sense that it always returns the same thing. (Indeed, as we will discuss shortly, this behavior is part-and-parcel of what it means to be a monad.) For instance, sampling from $\boxed{\llbracket \mathcal{J}o \rrbracket} : \mathbb{P}e$ will always result in $\llbracket \mathcal{J}o \rrbracket : e$. In the parlance of probability theory, such programs describe *degenerate distributions*.

3.1.3 The semantic value of *tall* as a probabilistic program

To model gradable adjectives like *tall*, Grove and Bernardy assume that $\llbracket tall \rrbracket$ is a probabilistic program of type $\mathbb{P}(e \rightarrow t)$. Their analysis uses the two monadic operators described above to model the interpretation of such adjectives in terms of probabilistic programs like the following one:

$$\begin{array}{l} d \sim \text{thresholdPrior} \\ \boxed{\lambda x.\text{height}(x) \geq d} \end{array}$$

This program first samples a random degree value $d : r$, where r is the type of real numbers, from *thresholdPrior*—a program of type $\mathbb{P}r$ —and then uses it inside the program $\boxed{\lambda x.\text{height}(x) \geq d}$ of type $\mathbb{P}(e \rightarrow t)$, thus providing a function of type $e \rightarrow t$ which depends on a probability distribution over degrees of height.

$$\begin{array}{lll}
\textit{Left identity} & \textit{Right identity} & \textit{Associativity} \\
x \sim \boxed{v} = k(v) & x \sim m = \boxed{x} & y \sim \left(\begin{array}{l} x \sim m \\ n(x) \end{array} \right) = \begin{array}{l} x \sim m \\ y \sim n(x) \\ o(y) \end{array}
\end{array}$$

Figure 4: The monad laws

Importantly, `thresholdPrior` can be anything, as long as it is of the right type (`Pr`). Its main function is to represent the constraints that the context—including comprehenders’ prior beliefs—imposes on d . For instance, one could assume that d is normally distributed with some mean μ and standard deviation σ , in which case the meaning of *tall* would be:

$$\begin{array}{l}
d \sim \mathcal{N}(\mu, \sigma) \\
\boxed{\lambda x. \text{height}(x) \geq d}
\end{array}$$

Under this assumption, the height threshold is sampled from—that is, *bound by*—the program $\mathcal{N}(\mu, \sigma) : \text{Pr}$ that computes a normal distribution.

3.1.4 Why it matters that \mathbb{P} is a monad

Because \mathbb{P} , together with $\boxed{\cdot}$ and (\sim) , is assumed to be a monad, it must satisfy the laws in Figure 4. Among these laws, Left identity guarantees that transforming a value v via $\boxed{\cdot}$ creates a “pure” probabilistic program that just returns v ; that is, v is the only value which may be sampled. Right identity guarantees that returning a value randomly sampled from m is just the same as computing a value from m . Associativity provides a syntactic convenience by allowing probabilistic programs to be re-bracketed: if one samples y from a complex probabilistic program that contains a use of (\sim) , one may also pull out the parts composing the program and, instead sample y from the last one. Together, the laws ensure that \mathbb{P} never tampers with the underlying compositional semantics (see Charlow 2014; Charlow 2020 for extensive discussion).

3.1.5 Extracting probabilities from probabilistic programs

Because Grove and Bernardy deal primarily with sentences containing vague predicates, they require not only a way of describing how probabilistic programs may be constructed and sampled from, but also a way of computing the probability of a particular value—e.g., the probability that a sentence containing some vague predicate is true. Thus, they require a method of going from programs m of type $\mathbb{P}\alpha$ to values of type r (real values). To satisfy this requirement, they (at least implicitly) use an *expected value* operator:

$$\mathbb{E}_{(\cdot)} : \mathbb{P}\alpha \rightarrow (\alpha \rightarrow r) \rightarrow r$$

Given a function f from values of type α to real numbers, $\mathbb{E}_{x \sim m} [f(x)]$ is the expected value of f , given the probability distribution over values of type α represented by m .⁹ If m returns

⁹An expected value of a function f is effectively an average over values $f(x)$ in that function’s range, weighted by the probability associated with x (in this case, as implied by the probabilistic program m).

truth values—i.e., if it is of type Pt —it can be associated with a probability by taking the expected value of the indicator function $\mathbb{1} : t \rightarrow r$, which maps \top (‘true’) to 1 and \perp (‘false’) to 0:

$$\begin{aligned} \mathbb{P} &: Pt \rightarrow r \\ \mathbb{P}(m) &= \mathbb{E}_{\tau \sim m} [\mathbb{1}(\tau)] \end{aligned}$$

For illustration, suppose we want to find the probability that the sentence *Jo is tall* is true. Taking the denotation of this sentence to be

$$\begin{aligned} \llbracket \textit{Jo is tall} \rrbracket &: Pt \\ \llbracket \textit{Jo is tall} \rrbracket &= d \sim \mathcal{N}(\mu, \sigma) \\ &\quad \text{height}(j) \geq d \end{aligned}$$

we use the probability operator \mathbb{P} to compute the probability

$$\begin{aligned} \mathbb{P} \left(\begin{array}{c} d \sim \mathcal{N}(\mu, \sigma) \\ \text{height}(j) \geq d \end{array} \right) &= \mathbb{E}_{\tau \sim \left(\begin{array}{c} d \sim \mathcal{N}(\mu, \sigma) \\ \text{height}(j) \geq d \end{array} \right)} [\mathbb{1}(\tau)] \\ &= \mathbb{E}_{d \sim \mathcal{N}(\mu, \sigma)} [\mathbb{1}(\text{height}(j) \geq d)] \end{aligned}$$

Thus the probability that *Jo is tall* is true is equal to the probability that $\text{height}(j) \geq d$, where d is a normally distributed random variable with mean μ and standard deviation σ .

3.1.6 Contexts in a probabilistic semantics

To model clause-embedding predicates, we need some way of representing the denotations of declarative clauses, which are standardly taken to be propositions. Following Grove and Bernardy (2023), we encode such representations by allowing the meanings of expressions to depend on *contexts*. Contexts, in our setting, are finite tuples of parameters that determine the semantic values of expressions. Thus, they are akin to models, possible worlds, or situations (see von Stechow and Heim 2021 and references therein). In addition to providing parameters that determine the denotations of expressions, contexts provide values for contextual parameters—e.g., the height threshold relevant to evaluating the meaning of a gradable adjective like *tall*. Taking κ to be the type of contexts (i.e., κ is an n -ary product, for some n), we may use the following notation to provide a new meaning for *tall*:

$$\begin{aligned} \llbracket \textit{tall} \rrbracket &: e \rightarrow \kappa \rightarrow t \\ \llbracket \textit{tall} \rrbracket &= \lambda x, c. \text{height}(c)(x) \geq d_{\textit{tall}}(c) \end{aligned}$$

$\text{height}(c)$ selects whichever component of c maps individuals to their heights, and $d_{\textit{tall}}(c)$ selects whichever component of c provides the contextual degree threshold relevant to determining the truth of the gradable adjective *tall*. In addition to settling facts about how the world is—e.g., people’s heights—contexts settle matters of vagueness and metalinguistic uncertainty—e.g., how tall one must be in order to be considered tall—as well as, possibly,

whether or not subjective predicates, like *tasty*, are true or false of some entity. Thus they may also be seen as akin to the “counterstances” of Kennedy and Willer (2016) and Kennedy and Willer (2022) or the “outlooks” of Coppock (2018).

Propositions in the current setting can now be conveniently viewed as sets of contexts, or functions of type $\kappa \rightarrow t$. Furthermore, following Grove and Bernardy (2023), the common ground may be viewed as a *distribution* over contexts, or a probabilistic program of type $P\kappa$. To update the common ground with a proposition, we make use of a function *observe*, which is defined, in turn, using a more primitive operation *factor*, whose role is to scale the distribution represented by the probabilistic program which follows it by some scalar value:¹⁰

$$\begin{aligned} \text{factor} &: r \rightarrow P\diamond \\ \text{observe} &: t \rightarrow P\diamond \\ \text{observe}(\phi) &= \text{factor}(\mathbb{1}(\phi)) \end{aligned}$$

\diamond is the unit type: it is inhabited by a single value—the 0-tuple (also written ‘ \diamond ’). It therefore carries no interesting information, implying that the role of *factor* is only to contribute a probabilistic effect, i.e., without computing a value.

Given a common ground $cg : P\kappa$, one can update it with the proposition $\phi : \kappa \rightarrow t$ by turning ϕ from a static into a dynamic proposition:

$$\begin{aligned} \text{update} &: (\kappa \rightarrow t) \rightarrow P\kappa \rightarrow P\kappa \\ \text{update}(\phi)(cg) &= c \sim cg \\ &\quad \text{observe}(\phi(c)) \\ &\quad \boxed{\kappa} \end{aligned}$$

Dynamizing propositions is thus a matter of *observing* them in the context provided by the relevant common ground.

To foreshadow our analyses a bit, each of the models we consider in this paper provides a representation of the common ground: at their heart, our models characterize distributions over contexts. The ways in which they differ from one another has to do with how the distributions over certain relevant parameters of a given context are evaluated, and in turn, how these distributions contribute to the predicted behavior of someone who makes an inference.

3.2 Our contribution: two levels of uncertainty

Our main contribution comes in how we model the common ground. Rather than representing the common ground as a probability distribution over contexts—i.e., as a program of type $P\kappa$ —we represent it as a probability distribution *over* probability distributions over contexts—i.e., as a program of type $P(P\kappa)$. By invoking the map P twice, we are effectively providing two layers, or levels, of probabilistic uncertainty.

¹⁰In the continuation-based setting of Grove and Bernardy 2023, *factor* is defined as

$$\text{factor}(x) = \lambda k.x * k(\diamond)$$

so that it scales its continuation by the relevant factor. For current purposes, we maintain a relatively abstract interface so that our main points aren’t obscured by implementation details.

$$\begin{array}{ll}
 \textit{Identity} & \textit{Composition} \\
 id^{\Downarrow} = id & (f \circ g)^{\Downarrow} = f^{\Downarrow} \circ g^{\Downarrow}
 \end{array}$$

Figure 5: The functor laws

We use the “inner” P to represent the uncertainty that is manifest on particular occasions of use and interpretation. Such uncertainty may, in principle, arise because of linguistic expressions which are vague or subjective, or it may be uncertainty related to beliefs that people have about the world. As an umbrella term, we refer to any of these sources of uncertainty as *contextual* uncertainty.

We use the “outer” P to represent *metalinguistic* uncertainty. Although there may, in general, be uncertainty about the values of linguistic parameters that govern the meanings of expressions, by regulating them on the outer layer, we take those values to be fixed on particular occasions of language use and interpretation. Thus one may regard the outer P as providing a distribution over possible *kinds* of occasions of use and interpretation—that is, which fix the values of parameters which are metalinguistically uncertain—while the inner P may be considered to be residual uncertainty that arises on particular occasions of use and interpretation, once the relevant type of occasion is fixed.

Which phenomena should be tethered to which layer of uncertainty is, importantly, up for debate and should ultimately be settled empirically. Our attempt to study the source of the gradience induced by factive predicates aims to help resolve this question in one of its manifestations. Thus to sharpen the distinction between the two hypotheses of Section 2.2, we ask whether the uncertainty giving rise to gradience among judgments of presupposition projection is (a) uncertainty that is settled as the occasion of use is fixed, or (b) an inherent property of particular uses and interpretations, so that presuppositions might project gradiently.

We note two important properties of the layering described above. First, the composition of P with itself has a certain formal license: because P is a monad, it is also a *functor*. This means that it comes with an operation $(\cdot)^{\Downarrow}$ (*map*) allowing one to perform pure operations on the values returned by probabilistic programs, while keeping their probabilistic effects intact. $(\cdot)^{\Downarrow}$ may be defined in terms of the monadic (\cdot) and (\sim) , as follows:

$$\begin{aligned}
 (\cdot)^{\Downarrow} &: (\alpha \rightarrow \beta) \rightarrow P\alpha \rightarrow P\beta \\
 f^{\Downarrow} &= \lambda m. \ x \sim m \\
 &\quad \boxed{f(x)}
 \end{aligned}$$

The two laws regulating functors are given in Figure 5.¹¹

Crucially, functors are *composable*, meaning that we can take the composition of the functor P with itself to obtain the new functor $P(P\alpha)$, whose $(\cdot)^{\Downarrow}$ may be defined simply as $(\cdot)^{\Downarrow\Downarrow}$.¹² Old operations are easily recast in the current setting involving structured uncertainty, that

¹¹Note that both laws may be proved from the monad laws of Figure 4.

¹²Indeed, because P is a monad, it is not only a functor, but an *applicative functor* (McBride and Paterson 2008), meaning that it comes with an operation

$$(\otimes) : P(\alpha \rightarrow \beta) \rightarrow P\alpha \rightarrow P\beta$$

is, by *mapping them* onto operations on higher-order probabilistic programs. Updates to the common ground, for instance, may be presented as follows:

$$\begin{aligned} \text{update}_2 &: (\kappa \rightarrow t) \rightarrow \mathsf{P}(\mathsf{P}\kappa) \rightarrow \mathsf{P}(\mathsf{P}\kappa) \\ \text{update}_2(\phi) &= \text{update}(\phi)^\Downarrow \end{aligned}$$

The second property of note is that, because $\mathsf{P}(\mathsf{P}\alpha)$ is obtained as the composition of functors, it provides a tight constraint on the way information may flow from one level to another; the flow is unidirectional, going from the outer level that regulates metalinguistic uncertainty to the inner level that regulates contextual uncertainty. As a result, it is possible for contextual uncertainty to remain even after questions of metalinguistic uncertainty have been settled—e.g., whether a semantically ambiguous expression has one interpretation versus another. But by necessity, settling contextual uncertainty also settles metalinguistic uncertainty.

This asymmetry is motivated by the general behavior of the two sources of uncertainty being modeled. To illustrate this, say someone makes the utterance *Jo is tall* in a noisy environment, rendering it ambiguous between *Jo is tall* and *Jo is small*. Moreover, say that, from the interlocutor’s perspective, the probability that *Jo is tall* was uttered is 0.7 and the probability that *Jo is small* was uttered is 0.3. Then (setting aside our commitment to employing contexts, momentarily), the metalinguistically uncertain *Jo is -all* can be assigned the following interpretation:

$$\begin{aligned} \llbracket \text{Jo is -all} \rrbracket &: \mathsf{P}(Pt) \\ \llbracket \text{Jo is -all} \rrbracket &= \tau \sim \text{Bernoulli}(0.7) \\ &\left\{ \begin{array}{ll} \begin{array}{l} d \sim \mathcal{N}(\mu_t, \sigma_t) \\ \text{height}(j) \geq d \end{array} & \tau \\ \begin{array}{l} d \sim \mathcal{N}(\mu_s, \sigma_s) \\ \text{size}(j) \leq d \end{array} & \neg\tau \end{array} \right. \end{aligned}$$

According to this interpretation, the meaning of *Jo is -all* depends on the Bernoulli-distributed variable $\tau : t$. If τ is \top (which occurs with a probability of 0.7), then the interpretation is the returned program which encodes the meaning of *Jo is tall*; whereas, if τ is \perp (which occurs with a probability of 0.3), then it is the returned program which encodes the meaning of *Jo is small*. Crucially, once the value of the random variable τ , which represents the metalinguistic uncertainty about what was uttered, is settled, one obtains a meaning having contextual uncertainty, encoded by a normal distribution over degrees of height or size, respectively. Thus, the probabilistic effects encoding contextual uncertainty depend on those encoding metalinguistic uncertainty (about the value of τ , in particular). But the former cannot, in turn, influence the latter, simply because they are part of the program which is *returned*; any parameters introduced by such effects are not in scope early enough.

We now turn to an account of factivity within this two-layered probabilistic setting.

called ‘*sequential application*’, which can apply an effectful function to an effectful argument, in order to sequence the effects. Applicatives also enjoy composability (so that $\mathsf{P}(\mathsf{P}\alpha)$ is also applicative), but we suppress this fact in the discussion for now, since applicatives provide a somewhat more powerful interface than we require.

3.3 The meaning of factivity

In general, we assume that clause-selecting predicates entail the complement clauses they select with some probability.¹³ For example, we may represent the meaning of *know* as follows, where τ_{know} selects from the context c a truth value determining whether to instantiate the meaning of *know* with a factive or a non-factive meaning:

$$\begin{aligned} \llbracket know \rrbracket &: (\kappa \rightarrow t) \rightarrow e \rightarrow \kappa \rightarrow t \\ \llbracket know \rrbracket &= \lambda\phi, x, c. \begin{cases} \text{know}(\phi)(x)(c) \wedge \phi(c) & \tau_{know}(c) \\ \text{know}(\phi)(x)(c) & \neg\tau_{know}(c) \end{cases} \end{aligned}$$

And likewise for all clause-selecting predicates. Those verbs which are always factive will have the meaning $\lambda\phi, x, c. \text{verb}(\phi)(x)(c) \wedge \phi(c)$ with probability 1, and those verbs which are never factive will have the meaning verb with probability 1.

It is important to note that while the entry provided above for *know* may appear to render it semantically ambiguous, we stress that it does not. On our account, ambiguity is a potential cause of metalinguistic uncertainty, but not of contextual uncertainty (ambiguities are resolved in context). Thus whether the above entry for *know* renders it ambiguous versus, say, *vague* is a matter of how the parameter τ_{know} is regulated; that is, whether its distribution is determined by metalinguistic uncertainty or contextual uncertainty.

We should also point out that the kind of analysis we present here is *ad hoc* in the sense that it provides no explanation of factivity beyond the entailment properties associated with predicates' interpretations. This approach is sufficient for current purposes, since we do not aim to provide an explanation of factivity. Rather, we aim to discover properties of its behavior—i.e., whether the gradience it exhibits is a manifestation of contextual uncertainty (supporting the fundamental gradience hypothesis) or metalinguistic uncertainty (supporting the fundamental discreteness hypothesis).

To sharpen this point, note that we could just as easily analyze factivity as arising from a source external to the predicate by assuming, e.g., a complementizer that encodes factivity (Kiparsky and Kiparsky 1970 *et seq*).

$$\begin{aligned} \llbracket that \rrbracket &: (\kappa \rightarrow t) \rightarrow ((\kappa \rightarrow t) \rightarrow e \rightarrow \kappa \rightarrow t) \rightarrow e \rightarrow \kappa \rightarrow t \\ \llbracket that \rrbracket &= \lambda\phi, v, x, c. \begin{cases} v(\phi)(x)(c) \wedge \phi(c) & \tau_{that}(c) \\ v(\phi)(x)(c) & \neg\tau_{that}(c) \end{cases} \end{aligned}$$

Given such a complementizer, we need not assume that clause-embedding predicates, such as *know*, themselves give rise to factive interpretations.¹⁴

$$\begin{aligned} \llbracket know \rrbracket &: (\kappa \rightarrow t) \rightarrow e \rightarrow \kappa \rightarrow t \\ \llbracket know \rrbracket &= \text{know} \end{aligned}$$

¹³We do not distinguish between factivity and veridicality for current purposes. This approach bears a resemblance to the general approach in Simons 2007 and Simons, Tonhauser, et al. 2010.

¹⁴This implementation—in which the semantic value of the complementizer operates on the semantic value of the predicate—is analogous to some neo-Davidsonian approaches to the semantics of propositional attitude verbs and complementizers (see, e.g., White and Rawlins 2018a, fn. 6) and may be adapted to other such approaches that regard the embedded clause as an intersective modifier of eventualities (Elliott 2016; Elliott 2020).

What this sort of approach requires, in turn, is that contexts provide information about which predicate a given complementizer co-occurs with, which must be available in order for the probability of projection to be modulated by predicate type (see Gordon and Chafetz 1990; Trueswell, Tanenhaus, and Kello 1993; MacDonald, Pearlmutter, and Seidenberg 1994; Garnsey et al. 1997; Altmann and Kamide 1999, i.a.). Insofar as the context makes such information available, this implementation could be used to yield the same range of statistical models as an implementation which assumes that factivity is driven principally by the lexical semantics of predicates.

4 Modeling

To investigate the theories of factivity and world knowledge possible under the framework described in Section 3, we implement Bayesian models in the Stan programming language ({Stan Development Team} 2023) via the CmdStanR interface (Gabry and Češnovar 2023). We fit these models using Degen and Tonhauser’s (2021) experimental data and then compare them in terms of their expected log pointwise predictive densities (ELPDs) computed under the widely applicable information criterion (WAIC; Watanabe 2013; Gelman, Hwang, and Vehtari 2014) as implemented in R’s loo package (Vehtari et al. 2023). This measure quantifies how well each model fits the data, while also penalizing each for how complex it is—i.e., the effective number of parameters it uses to fit the data.

In Section 4.1, we formalize our assumptions about the link between higher-order probabilistic programs of the kind described in Section 3 and participants’ response behavior. We then describe each of the models of factivity possible under our framework in Section 4.2 before reporting our comparisons of these models given Degen and Tonhauser’s projection experiment data in Section 4.3.

4.1 Linking to response behavior

To connect the probabilistic programs associated with sentences to actual data, we need linking assumptions that relate the former to the inference judgments experimental participants report on a slider scale between ‘no’ and ‘yes’. Following Degen and Tonhauser (2022), we model these responses as a distribution of values on the unit interval.

We specify our models by defining a class of functions respond^σ , each of which takes a distribution over contexts m of type $\text{P}\kappa$ and a possible inference ϕ of type $\kappa \rightarrow t$ and associates it with a distribution over slider responses on the unit interval:

$$\begin{aligned} \text{respond}_{(\cdot)}^\sigma &: \text{P}\kappa \rightarrow (\kappa \rightarrow t) \rightarrow \text{Pr} \\ \text{respond}_{c \sim m}^\sigma(\phi(c)) &= \mathcal{N}(x, \sigma) \text{ T}[0, 1] \\ &\text{ where } x = \mathbb{P} \left(\begin{array}{c} c \sim m \\ \boxed{\phi(c)} \end{array} \right) \end{aligned}$$

In essence, this implementation of response behavior assumes that participants compute the probability x of the inference ϕ by determining whether or not the inference is true in a context, weighted by how likely that context is under the probabilistic program m . They

then attempt to respond with x , but due to factors independent of the process by which x is computed—e.g., inaccuracies in their ability to perfectly target x on the response scale—they produce an actual response that is normally distributed (with standard deviation σ) around x , truncated to $[0, 1]$.¹⁵

To give a schematic example, let's say that the common ground of interest is characterized by a probabilistic program `commonGround` of type $P(P\kappa)$. Then, the following program of type Pr characterizes the distribution of slider responses on the unit interval, where a response reflects a judgment of certainty about the truth of the sentence *Grace visited her sister*, given the information *Susan knows that Grace visited her sister*:

$$\begin{aligned}
 & m \sim \text{commonGround} \\
 & \text{respond}_{c \sim m'}^{\sigma} (\llbracket \text{Grace visited her sister} \rrbracket^c) \\
 & \quad \text{where } m' = \text{update}(\llbracket \text{Susan knows that Grace visited her sister} \rrbracket)(m) \\
 = & m \sim \text{commonGround} \\
 & \mathcal{N}(x, \sigma) \top[0, 1] \\
 & \quad \text{where } x = \mathbb{P} \left(\begin{array}{l} c \sim m \\ \text{observe}(\llbracket \text{Susan knows that Grace visited her sister} \rrbracket^c) \\ \llbracket \text{Grace visited her sister} \rrbracket^c \end{array} \right)
 \end{aligned}$$

This probabilistic program first samples a distribution over contexts m of type $P\kappa$. Under this distribution, the parameters regulating metalinguistic knowledge have been *fixed*, but the parameters regulating contextual uncertainty still remain indeterminate. The program then computes a distribution over responses by doing a couple of things inside the scope of the \mathbb{P} operator: (i) it samples a context from m , which it uses to perform Bayesian update—by observing that *Susan knows that Grace visited her sister* is true—before (ii) returning \top or \perp , depending on which is the interpretation of *Grace visited her sister*, given that context.

For the purpose of specifying our models abstractly, we assume the type κ of contexts to be a product $t^m \times t^n$, where the inhabitants of t^m are m -tuples of truth values τ_v determining whether or not the complement of each predicate under consideration indeed *projects*, and the inhabitants of t^n are n -tuples of truth values τ_w determining whether or not each fact under consideration related to world knowledge is true or false. Thus, each model will be of the form

$$\text{commonGround} : P(P(t^m \times t^n))$$

¹⁵The likelihood assumed here is known as a truncated normal distribution, which we express using STAN's notational convention for truncation ($\top[0, 1]$) throughout. This assumption is analogous to the one that Degen and Tonhauser (2021) make in using a linear mixed model—though, in using a *truncated* normal, we additionally capture the boundedness of the response scale. An alternative likelihood sometimes used with bounded response scales is a Beta distribution (see, e.g., Degen and Tonhauser 2022). This assumption is not strictly appropriate for bounded response scales that include their endpoints—e.g., a response scale on the closed interval $[0, 1]$ rather than the open interval $(0, 1)$ —because Beta distributions only have support on the open interval—i.e., they exclude $\{0, 1\}$ (see Liu and Eugenio 2018 and references therein). It is particularly problematic in the current context, where endpoint responses are meaningful by hypothesis. Truncated normals do not have this problem because they can have support on the closed interval. Zero-one inflated Beta distributions are another option (again, see Liu and Eugenio 2018), but they bring their own conceptual challenges in the current context because they effectively require the assumption that all models assume some amount of discreteness.

Updating any such representation of the common ground with a proposition and predicting the distribution of judgments generated by an inference is a matter of following the procedure outlined above. Using the same example, we would obtain the following characterization of this distribution:

$$\begin{aligned}
 & m \sim \text{commonGround} \\
 & \mathcal{N}(x, \sigma) \top[0, 1] \\
 & \text{where } x = \mathbb{P} \left(\begin{array}{l} \langle \tau_v, \tau_w \rangle \sim m \\ \text{observe}(\llbracket \textit{Susan knows that Grace visited her sister} \rrbracket^{\langle \tau_v, \tau_w \rangle}) \\ \llbracket \textit{Grace visited her sister} \rrbracket^{\langle \tau_v, \tau_w \rangle} \end{array} \right)
 \end{aligned}$$

If we take $\text{know}(\tau_v)$ to be the component of τ_v that says whether or not the complement of *know* projects, and $\text{grace}(\tau_w)$ to be the component of τ_w that says whether or not Grace visited her sister, then the program above can be rephrased as follows:

$$\begin{aligned}
 & m \sim \text{commonGround} \\
 & \mathcal{N}(x, \sigma) \top[0, 1] \\
 & \text{where } x = \mathbb{P} \left(\begin{array}{l} \langle \tau_v, \tau_w \rangle \sim m \\ \text{know}(\tau_v) \vee \text{grace}(\tau_w) \end{array} \right)
 \end{aligned}$$

That is, because all that is required for *Grace visited her sister* to be entailed by the common ground is for the complement of *know* to project (i.e., for $\text{know}(\tau_v)$ to be \top) or for it to be entailed by prior knowledge (i.e., for $\text{grace}(\tau_w)$ to be \top), its semantic value is equivalent to a disjunction.¹⁶

The general set-up described here will remain invariant under the theories considered in the rest of this section. What varies is the structure of `commonGround`, and, crucially, whether the parameters regulating world knowledge and factivity are understood as being governed by metalinguistic uncertainty or contextual uncertainty.

4.2 Models of factivity

We now provide our four models of factivity and prior world knowledge, which we fit to Degen and Tonhauser’s projection experiment data.¹⁷ Each of these models assumes that factivity is either discrete or gradient and that world knowledge is either discrete or gradient in their contributions to inference. Indeed, the models we present here make these assumptions wholesale, either regarding every predicate as making a discrete contribution or regarding every predicate as making a gradient contribution (and likewise for world knowledge). Our current purpose is to compare these strong versions of the two hypotheses stated in Section 2, since we take it that prior work assumes that either one or the other is true of all predicates. Nonetheless, the possibility that predicates vary with respect to whether their factivity is discrete or gradient is an interesting one. We investigate this question in ongoing work.

¹⁶The full model specifications presented in Appendix A make crucial use of this fact.

¹⁷See Appendix A for the full model specifications.

4.2.1 Factivity as a fundamentally discrete phenomenon

The first theory we consider is consistent with the classical view of factivity, implicit in Kiparsky and Kiparsky (1970) and Karttunen (1974), *inter alia*. It regards clause-embedding predicates as either triggering or not triggering factive inferences on particular occasions of use, depending on their interpretations. The uncertainty about whether or not a given predicate’s complement clause projects is thus regarded as metalinguistic in nature.

We take it that most classical accounts of factivity do not take a stance on how world knowledge might affect these inferences. There are two possibilities implied by our framework, which we discuss in turn.

The first possibility allows for uncertainty related to world knowledge to manifest itself as contextual uncertainty, which may, in turn, make individual judgments of truth uncertain. We refer to this theory as the *discrete-factivity* theory, emphasizing that it regards the contribution factivity makes as fundamentally discrete in nature. It gives rise to the following common ground:

$$\begin{aligned}
 &\text{discrete-factivity} : P(P\kappa) \\
 \text{discrete-factivity} &= \langle v, w \rangle \sim \text{priors} \\
 &\quad \tau_v \sim \text{Bernoulli}(v) \\
 &\quad \tau_w \sim \text{Bernoulli}(w) \\
 &\quad \langle \tau_v, \tau_w \rangle
 \end{aligned}$$

The aspect of this model crucial to the way in which it regards factivity is the location of the sampling statement ‘ $\tau_v \sim \text{Bernoulli}(v)$ ’; in particular, it is crucial that this statement occurs *prior* to returning the probabilistic program of type $P\kappa$ that characterizes contextual uncertainty—that is, outside of the orange boxes. As a result, whether or not the complement of a given predicate projects is fixed in individual utterance contexts. By contrast, the sampling statement ‘ $\tau_w \sim \text{Bernoulli}(w)$ ’ is part of the returned program, rendering world knowledge contextually uncertain.

The second possibility understands both factivity and world knowledge as discrete in nature.

$$\begin{aligned}
 &\text{wholly-discrete} : P(P\kappa) \\
 \text{wholly-discrete} &= \langle v, w \rangle \sim \text{priors} \\
 &\quad \tau_v \sim \text{Bernoulli}(v) \\
 &\quad \tau_w \sim \text{Bernoulli}(w) \\
 &\quad \langle \tau_v, \tau_w \rangle
 \end{aligned}$$

This *wholly-discrete* version of the theory, which we do not take anyone to endorse but whose possibility is implied by our framework, effectively hypothesizes that no inferences display uncertainty in context: under this version, the sampling statement ‘ $\tau_w \sim \text{Bernoulli}(w)$ ’ has also been moved outside of the outer orange box. Any gradience displayed in people’s measured inferences must therefore be due to response error.

4.2.2 Factivity as a fundamentally gradient phenomenon

The theory which regards factivity as fundamentally gradient in nature does so by pushing what would otherwise be metalinguistic uncertainty about factivity onto the contextual uncertainty level. This theory also has two possible implementations that vary with respect to whether world knowledge is modeled as contextually uncertain or metalinguistically uncertain. We refer to the version of the theory that regards both factivity and world knowledge as contextually uncertain as the *wholly-gradient* version. We take this version to be the one that comes closest to implementing Tonhauser, Beaver, and Degen’s (2018) proposal among the four we consider here.

We obtain the corresponding model by making a small modification to the discrete-factivity model—that is, by changing the location of the relevant sampling statement:

$$\begin{aligned} \text{wholly-gradient} &: P(\mathcal{P}\kappa) \\ \text{wholly-gradient} &= \langle \mathbf{v}, \mathbf{w} \rangle \sim \text{priors} \\ &\quad \tau_{\mathbf{v}} \sim \text{Bernoulli}(\mathbf{v}) \\ &\quad \tau_{\mathbf{w}} \sim \text{Bernoulli}(\mathbf{w}) \\ &\quad \langle \tau_{\mathbf{v}}, \tau_{\mathbf{w}} \rangle \end{aligned}$$

We refer to the alternative that regards world knowledge as discrete and factivity as gradient as the *discrete-world* version. As with the wholly-discrete model described above, we do not take anyone to endorse version of the theory, but we include it, since it is a possibility implied by our framework. Under this version, the locations of the sampling statements which were used to encode the discrete-factivity model are switched:

$$\begin{aligned} \text{discrete-world} &: P(\mathcal{P}\kappa) \\ \text{discrete-world} &= \langle \mathbf{v}, \mathbf{w} \rangle \sim \text{priors} \\ &\quad \tau_{\mathbf{w}} \sim \text{Bernoulli}(\mathbf{w}) \\ &\quad \tau_{\mathbf{v}} \sim \text{Bernoulli}(\mathbf{v}) \\ &\quad \langle \tau_{\mathbf{v}}, \tau_{\mathbf{w}} \rangle \end{aligned}$$

4.3 Comparisons

To construct our models, we use a *pipelined* approach.¹⁸ Following this approach, we first fit a model of Degen and Tonhauser’s norming data (experiment 2a) in order to obtain posterior distributions for parameters associated with the forty pairs of complement clauses and facts, which we refer to as *contexts*.¹⁹ We then use these posterior distributions as prior distributions for the corresponding parameters in the four models of Degen and Tonhauser’s projection experiment data (experiment 2b) implied by the framework described in Section 3.

Figure 6 (left plot) provides ELPDs estimated for the four models, based on log-likelihoods computed from Degen and Tonhauser’s experimental data. We observe that the discrete-factivity model captures the data the best, while the wholly-discrete model trails behind it;

¹⁸We provide additional formal details about the modeling pipeline in Appendix A.

¹⁹Note that we are overloading the term *context* here. This notion of context is not the same as our formal notion.

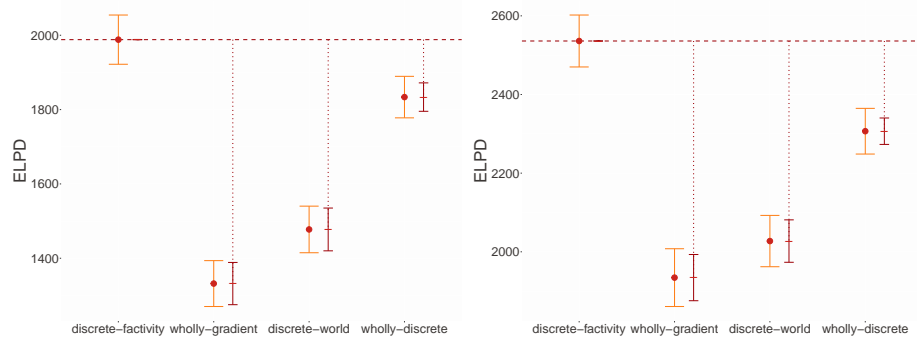


Figure 6: Left: ELPDs for the four models. Right: ELPDs for the four model evaluations on our replication experiment data. Dotted lines indicate estimated differences between each model and the discrete-factivity model. Error bars indicate standard errors.

meanwhile, the wholly-gradient and discrete-world models—the two that assume factivity to be gradient—perform the worst.

Thus, we have preliminary evidence that the best model of Degen and Tonhauser’s data treats factive presupposition projection as a discrete phenomenon and the inferences contributed by world knowledge as gradient. Meanwhile, by simply modifying the discrete-factivity model so that it treats factivity as gradient, one goes from the *best*-performing model to the *worst*-performing one.

To give a sense of the performance of the models as assessed against the inference judgment data, the posterior predictive distributions for each model are plotted for six predicates, for all contexts combined, in Figure 7 (see Figure 16 of Appendix B.2 for all predicates). These plots provide a visual indicator of how well each model fits the distribution of judgments for these predicates across contexts: the closer a particular curve is to the shape of the histogram, the better the corresponding model fits the data.

We observe here that the models that assume that either factivity or world knowledge is discrete are better able to capture dips in the frequency of responses in the middle of the scale than the wholly-gradient model. As one might expect, the wholly-gradient model predicts distributions that are much smoother than the models that assume some form of discreteness. This behavior is the main reason the wholly-gradient model fits the inference judgment data the worst of any model.

One reason the models assuming some amount of discreteness can capture the dips in frequency toward the middle of the scale is that they effectively model the distribution over inference judgments as a mixture of distributions: at least one with a mode at 1 and another with a mode determined by the structure of the particular model. In contrast, the wholly-gradient model only assumes a single distribution.²⁰ This property of the models that assume some amount of discreteness makes them more complex than the wholly-gradient model—in

²⁰Multi-modality may still arise in the posterior predictive distributions for the wholly-gradient model; see, e.g., the posterior predictive distribution for *confirm* in Figure 7. Such multi-modality can only arise due to participant random effects. See Appendix A for formal details of these random effects.

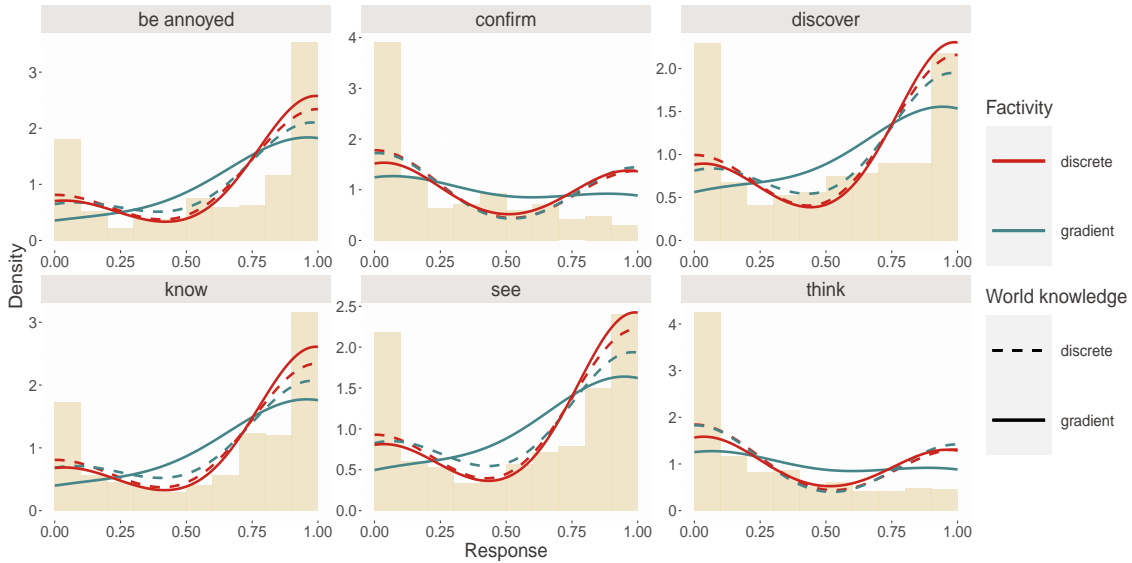


Figure 7: Posterior predictive distributions (with simulated participant intercepts) of all four models for six predicates from Degen and Tonhauser’s (2021) projection experiment, for all contexts combined. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

the sense that they have more effective parameters—but as the pattern of ELPDs in Figure 6 shows, this additional complexity is offset by how much better they fit the data. (Recall that ELPD explicitly quantifies how well a model fits the data while penalizing model complexity as measured by the effective number of parameters.)

Notably, none of the four models fits the data perfectly. For instance, the canonically non-projective predicates *think* and *pretend* have distributions which all four of our models appear to have difficulty capturing, at least by visual inspection of Figure 16. This difficulty appears to be due to an anti-veridicality inference associated with these predicates—i.e., that the complement clause is *not* true.²¹

None of our models naturally captures such anti-veridicality inferences, since we consider denotations that vary only with respect to veridicality and non-veridicality—not anti-veridicality. Assuming a three-way distinction among veridicality, non-veridicality, and anti-veridicality would effectively allow the wholly-discrete, discrete-factivity, and discrete-world models to mix in another distribution with a mode at 0, thereby allowing them to fit the data better. Even so, such a modification is unlikely to change the fact that the wholly-gradient model performs the worst, which is apparently due to its inability model the bimodality in the response distributions as well as the other models do. Whether or not it will can only be

²¹These inferences likely arise from different sources: e.g., the lexical semantics of *pretend*, but a conversational implicature in the case of *think*. More generally, inferences of this kind may, in principle, arise for *any* predicate on its non-factive interpretation (assuming factivity is discrete), which may shed additional light on the mass of responses at or close to 0 in the empirical data across predicates.

answered by implementing the relevant models, however, which we leave for future work.

5 Evaluations

Strictly speaking, the comparisons we report in Section 4 are *post hoc*; and while the results are suggestive, we cannot draw firm conclusions from these model comparisons without a replication of Degen and Tonhauser’s experiment. In Section 5.1, we report such a replication, finding the same pattern of model comparison results: the models that assume that factivity is discrete reliably outperform the models that assume it is gradient. To ensure that these results are not somehow driven by the particular discourse contexts used in Degen and Tonhauser’s experiments, we collect two additional datasets that use the same paradigm but that mask the contents of the embedded clause in two distinct ways. In these additional experiments, which we report in Section 5.2, we again find that the discrete-factivity models outperform all of the other models.

5.1 Experiment 1: held-out projection experiment data

5.1.1 Materials

Our materials and methods were identical to those of Degen and Tonhauser (2021).

5.1.2 Participants

We collected data from 300 participants using Amazon Mechanical Turk, paying each participant one dollar. Each participant was required to pass the qualification test described in White, Hacquard, and Lidz 2018, in order to ensure that they were a native speaker of American English. We removed data from two participants who claimed to have technical difficulties completing the experiment, and from ten more whose performance was more than two standard deviations below the mean on the six control items, leaving us with data from a total of 288 participants.

5.1.3 Results

Figure 8 shows the item means from our replication experiment plotted against Degen and Tonhauser’s original experiment, along with the means obtained by collapsing across contexts. We observe that there is strong agreement (Spearman’s $r = 0.98^{***}$) between these means and those obtained from Degen and Tonhauser’s data.

5.1.4 Model fitting

To evaluate the four models using this data, we obtain, from each model, means μ_v and standard deviations σ_v of the marginal posterior distribution of the log-odds of projection for each predicate, as well as means μ_ω and standard deviations σ_ω of the marginal posterior distribution of the log-odds certainty for each context. We then use normal distributions with

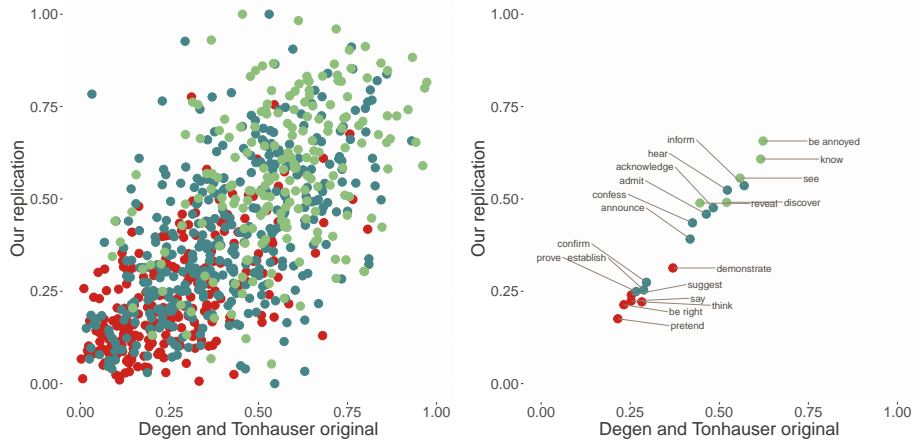


Figure 8: Degen and Tonhauser’s (2021) projection data versus our replication. Left: item means (Spearman’s $r = 0.68^{***}$). Right: verb means (Spearman’s $r = 0.98^{***}$). For both, “non-factive” verbs are in red, “optionally factive” verbs are in teal, and “canonically factive” verbs are in green.

these means and standard deviations as prior distributions for the corresponding parameters in the models constructed for the evaluations.²²

5.1.5 Model comparison

Figure 6 (right plot) provides ELPDs estimated for the four model evaluations, based on log-likelihoods computed from the data obtained in our replication experiment. The pattern of goodness-of-fit observed here is extremely close to that exhibited by the original model comparison on the left in Figure 6: the discrete-factivity model fares the best, followed by the wholly-discrete model. The wholly-gradient and discrete-world models fare the worst.

Thus we have evidence that the differences in performance among these models that was reported in the previous section are quite robust, at least when assessed using data from Degen and Tonhauser’s experimental task. The next two experiments provide a test of the models in a somewhat different setting—one in which the uncertainty contributed by the linguistic contexts of the predicates of interest is sent to an extreme.

5.2 Experiments 2 and 3: non-contentful contexts

We now test the inferred posterior distributions over probabilities of projection on held-out data from two experiments in which the context of each predicate is stripped of the rich lexical content that partially governs the inferences produced in the original experiment. We obtain contexts for this evaluation in two ways. In our Experiment 2, we *bleach* each predicate’s complement clause so that it is just *a particular thing happened*. In the Experiment 3, we use a *templatic* complement clause of the form *X happened*.

²²See Appendix A.3 for further details concerning these models.

These manipulations serve two purposes. First, they allow us to assess the performance of the four models when the source of variance among inference judgments contributed by prior world knowledge is removed. Second, they put the predicates in environments in which knowledge about the context is minimal; as a result, they may produce greater uncertainty in people's inferences. This additional uncertainty could confer an *a priori* advantage to the wholly-gradient model, which considers all inferences, even those triggered by projective predicates, to be beset with some uncertainty. We thus consider these evaluations to be a stronger test of the discrete-factivity model's edge over the wholly-gradient model than the original experiments were.

5.2.1 Materials

To construct the bleached items, each of the twenty predicates investigated in Degen and Tonhauser's experiment was placed in a context in which its subject was one of the proper names from the original experiment, and in which its complement clause was just *a particular thing happened*. On each trial, participants were provided with a background context that was intended to make the prompt as natural as possible. The only thing that varied in this background context from one trial to the next was the name of the individual who makes the relevant utterance. Finally (taking that individual's name to be *P*), participants were prompted to answer the question *Is P certain that that thing happened?* on a sliding scale with 'no' on the left and 'yes' on the right. The following experimental trial, for example, involves the predicate *pretend*:

You are at a party. You walk into the kitchen and overhear Linda ask somebody else a question. Linda doesn't know you and wants to be secretive, so speaks in somewhat coded language.

Linda asks: "*Did Tim pretend that a particular thing happened?*"

Is Linda certain that that thing happened?

no yes

Next

In addition to the twenty bleached items, participants saw six control items which were constructed in order to somehow incorporate a bleached subordinate clause; for example, *Did Madison have a baby, despite the fact that a particular thing happened?*. All six control items had an intended response of 1.

5.2.2 Experiment 3: templatic items

To construct the templatic items, each of the same twenty predicates was placed in a context in which its subject was, again, a proper name from the original Degen and Tonhauser ex-

periment, and in which its complement clause was *X happened*. A background context was provided on each trial, so that the prompt was natural. Background contexts, again, only differed from one another in the name of the individual who makes the relevant utterance. Given a trial on which the individual *P* was the speaker, participants were prompted with the question *Is P certain that X happened?*, which they answered on a sliding scale with ‘no’ on the left and ‘yes’ on the right. The following example trial features the predicate *pretend*:

You are at a party. You walk into the kitchen and overhear William ask somebody else a question. The party is very noisy, and you only hear part of what is said. The part you don't hear is represented by the 'X'.

William asks: "Did Ray pretend that X happened?"

Is William certain that X happened?

no yes

Participants, again, saw six control items which were constructed in order to incorporate a templatic subordinate clause; for example, *Did Madison have a baby, despite the fact that X happened?*.

5.2.3 Participants

For each experiment, we collect data from 50 new participants using Amazon Mechanical Turk, paying each participant one dollar. Each participant was, again, required to pass the qualification test described in White, Hacquard, and Lidz 2018, and any participant whose average score on the control items did not fall within two standard deviations below the mean of all participant's responses was excluded from the analysis. Using this criterion, three participants' data was excluded from Experiment 2, leaving 47 participants for analysis; and one participant was excluded from Experiment 3, leaving 49 participants for analysis.

5.2.4 Results

We observe in Figure 9 that the responses elicited by both the bleached (Spearman's $r = 0.97^{***}$) and templatic (Spearman's $r = 0.87^{***}$) items track the gradient knowledge about factivity that people deploy in the typical contentful setting extremely well. Not only is the same type of gradience observed among predicates when they are placed in bleached or templatic contexts, but the rankings among predicates are maintained almost entirely.²³ This

²³Notably, the range of average ratings for predicates in Experiment 3 is not as wide as exhibited in the previous experiments, with most falling between 0.5 and 0.75, suggesting that there may have been a great deal of uncertainty governing the inferences produced from this task.

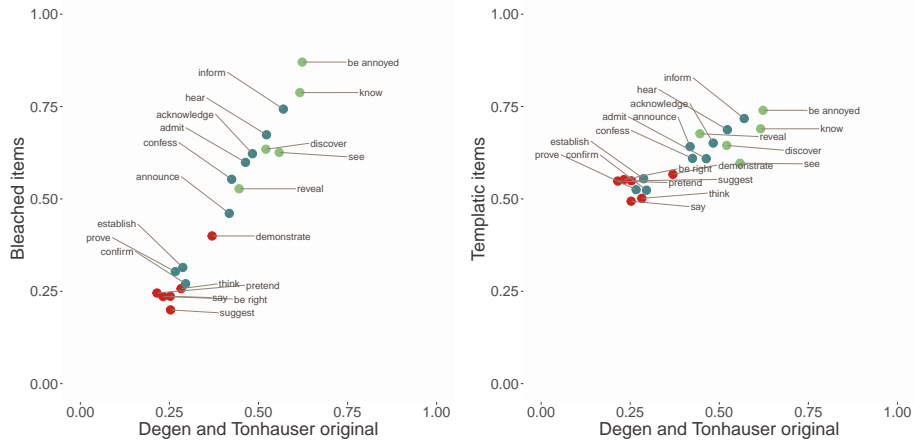


Figure 9: Degen and Tonhauser’s (2021) projection data versus data from contexts with minimal lexical content. Left: bleached data (Spearman’s $r = 0.97^{***}$). Right: templatic data (Spearman’s $r = 0.87^{***}$). For both, “non-factive” verbs are in red, “optionally factive” verbs are in teal, and “canonically factive” verbs are in green.

finding furthermore suggests that it is safe to compare modeling results obtained from only bleached or templatic items to those obtained from contentful items—e.g., the clustering results of Kane, Gantt, and White 2022 discussed in Section 2—at least if those results pertain to the aggregate responses for individual predicates.

5.2.5 Model fitting

To evaluate the four models using both the bleached and templatic data, we use the same means μ_v and standard deviations σ_v of the marginal posterior distributions of the log-odds of projection that we used for the evaluations on the replication experiment data. As before, we use normal distributions with these means and standard deviations as prior distributions for the corresponding parameters in the models constructed for the evaluations. Then, in each evaluation, we infer a distribution over the parameters σ_ω and ω that regulate the certainty associated with either the bleached or the templatic context.²⁴

5.2.6 Model comparison

Figure 10 provides ELPDs for all four models evaluated on Experiments 2 and 3. We see, here, that the discrete-factivity model fares the best in both experiments, while the other three models fare comparably with each other. It is somewhat remarkable that the more fine-grained differences among models observable from both the original fits and the evaluation on the replication data do not appear to hold up under the current evaluation; for example, the wholly-discrete model no longer appears distinguished from the wholly-gradient and

²⁴See Appendix A.4 for further details concerning these models.

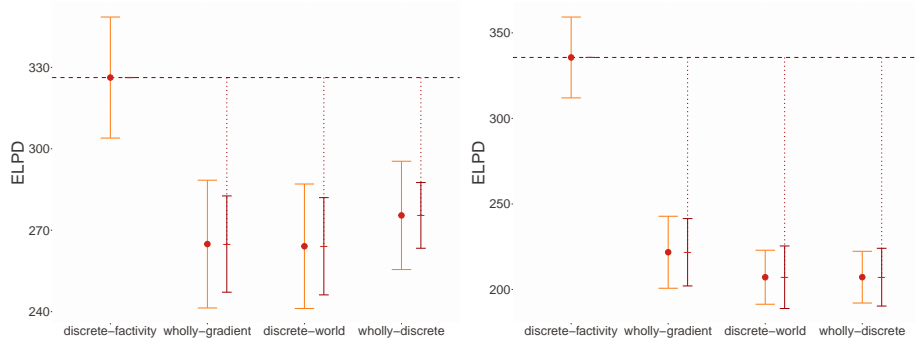


Figure 10: ELPDs for the four model evaluations on the bleached data (left) and the templatic data (right). Dotted lines indicate estimated differences between each model and the discrete-factivity model. Error bars indicate standard errors.

discrete-world models by its better performance. Rather, the discrete-factivity model seems to have a unique advantage.

This difference is especially notable for Experiment 3, given the somewhat squashed average responses seen across verbs in Figure 9. That is, despite the high amount of uncertainty which this task may have produced, such uncertainty seems to have been filtered through the discrete behavior of factive inferences. The uncertainty about such inferences appears to relate to the interpretation of a given predicate, rather than a gradient contribution which that predicate makes to an inference.

To give a sense of the differences among the model evaluations on the bleached and templatic data, Figures 17 and 18 of Appendix B.2 show the posterior predictive distributions of these evaluations for all predicates.

6 Consequences for theories of factivity

We have compared four models of the task reported in Degen and Tonhauser 2021, each fit to their experimental data. These models differ from one another along two axes: (a) whether they consider the contribution of prior world knowledge to inferences about the truth of the clause embedded by a predicate to be gradient or discrete; and (b) whether they consider the contribution of the relevant predicate itself to these inferences to be gradient or discrete. The contribution of a given factor to an inference is “gradient” if varying that factor produces a continuous effect on the magnitude of the judgment associated with the inference; and it is “discrete” if varying the factor affects the probability with which the inference is judged as certain, versus remains unaffected.

Our initial comparison of the four models found that the discrete-factivity model best accounts for the distributions of judgments in Degen and Tonhauser’s experimental data. That is, the model which regards the contribution of prior world knowledge to such inferences as gradient and the contribution of a given predicate to such inferences as discrete (as assessed by ELPDs, plotted in Figure 6, left). Moreover, follow-up evaluations of the four models confirmed the initial comparison: the same model best accounts for held-out data from a

replication of Degen and Tonhauser’s experiment, for which distributions over the parameters of interest are extracted from the posteriors of the initial models (Figure 6, right plot). The discrete-factivity model also best accounts for data from two tasks in which predicates are placed in contexts with minimal lexical content (Figure 10). Taken together, these results provide strong evidence that the observed gradience among the clause-embedding predicates studied by Degen and Tonhauser is *metalinguistic*. Different clause-embedding predicates differ in the frequencies with which they trigger projective inferences, but the contribution a predicate makes on particular occasions of use and interpretation is *discrete*, either producing the relevant inference or not producing it at all.

What do these conclusions mean for factivity as a property and the notion of ‘factive predicate’ as a class? Are there factive predicates? We propose that our results support an account of factivity whereon it is a semantically live property of expressions, but a property that may be observed on only some uses of those expressions. Indeed, many predicates which have traditionally been considered factive may, in fact, be systematically ambiguous. We take this finding to largely confirm what prior work stretching back to Karttunen 1971 has (at least implicitly) assumed in discussing ‘semifactive predicates’. These predicates may support factive readings in certain contexts, and with some proclivity which varies by individual predicate, or by class of predicate (Kane, Gantt, and White 2022). Hence, our results, in conjunction with Kane, Gantt, and White’s, are consistent with a fairly conservative picture of factivity, according to which it is an optional property of at least a subset of the predicates that Degen and Tonhauser investigate.

6.1 Factivity as an epiphenomenon

Our proposal is crucially a proposal about semantic properties of expressions. Yet, it is consistent with an account of factivity on which projection is intimately tied to properties of the discourse in which the expressions of interest are embedded.²⁵ A prime example of such an account can be found in Simons, Beaver, et al. 2017, who rely crucially on the notion of a question under discussion (QUD; Roberts 2012). On their account, whether or not the complement of a clause-embedding predicate projects varies according to prosodic and contextual factors associated with the QUD. Simons, Beaver, et al.’s main aim in giving this account is to do away with factivity as a semantic property of expressions, arguing rather that projective inferences are simply those which are backgrounded by the QUD, while non-projective inferences are those that are at-issue (and generally, entailed).

We do not believe that completely doing away with the notion of factivity as a semantic property will be possible for reasons that we have already mentioned in passing (Footnote 6): insofar as one is willing to countenance that discrete choices are made about the identity of the QUD on individual occasions of interpretation (*pace* Tonhauser, Beaver, and Degen 2018), one could posit that there is no indeterminacy in the interpretations for the relevant string and, rather, that there is uncertainty over possible QUDs against which the string is interpreted; but it is not clear how to reconcile such an account with the observation that (classes

²⁵See Qing, Goodman, and Lassiter 2016 for an approach along these lines within the Rational Speech-Act framework (Frank and Goodman 2012; Goodman and Stuhlmüller 2013).

of) predicates are associated with particular levels of gradience without saying that lexical knowledge somehow conditions QUD choice. Such lexical knowledge could be knowledge one has about the kinds of discourses in which a predicate or class of predicates is used—and therefore not semantic in nature—but this assumption raises a further question about why such knowledge would predict predicates’ syntactic distributions, as Kane, Gantt, and White (2022) show that it does. It seems much more plausible that this knowledge is at least partly semantic in nature.

So how could a semantic notion of factivity be integrated with accounts that intimately tie projection to properties of the discourse in which the expressions of interest are embedded? In answering this question, we believe it will be fruitful to combine standard dynamic accounts of presupposition projection (Heim 1992 *et seq*) with a probabilistic framework like Grove and Bernardy’s (2023) and our extension.

In broad strokes, dynamic accounts of presupposition projection in the Heimian tradition associate factive predicates with constraints on the contexts they can be used in—generally, requiring that the common ground entails the content of the factive predicate’s embedded clause. Such constraints could be stated in Grove and Bernardy’s framework (coupled with our extension) as constraints imposed by the predicate on the distribution over contexts associated with the common ground. Uncertainty about whether a predicate is factive or not thus implies uncertainty about what constraints to impose on that distribution over contexts.

Under such an account, properties of the QUD correlate with projection because possible QUDs are constrained by the common ground in at least the sense that the QUD cannot be trivial—i.e., the common ground cannot entail an answer to the QUD. Hence, insofar as a factive variant of a predicate is more probable, QUDs that are trivial under common grounds that are made more probable by that variant will have lower probability.

6.2 Which predicates are factive?

Which clause-embedding predicates *do*, in fact, belong in the class of factives is not a question whose answer we have formally pursued here. As we note in Section 2, the question we address in this paper is independent of whether or not there is a distinct class of factive predicates (or subclasses thereof). But to retrace the discussion of Section 2, we note that Kane, Gantt, and White (2022) have already done relevant work on this front by investigating how best to cluster the predicates of interest (and many other predicates) into semantic classes that are predictive of their syntactic distributions. Among other datasets, Kane, Gantt, and White rely on the MegaVeridicality dataset (White and Rawlins 2018b), which Degen and Tonhauser (2022) also use to support their findings of gradience among predicates’ veridicality inferences. Kane, Gantt, and White find that *emotive* predicates, such as *love* and *be pleased*, yield the most strongly factive inferences, and that such predicates are followed by *discourse commitment* predicates, which include *know*. Thus we are optimistic that clusterings of predicates based solely on diagnostics of factivity will also give rise to semantically potent classes.

In this vein, restricting attention only to the predicates that Degen and Tonhauser study yields a promising outlook. Figure 11 plots the posterior probability of projection associated with the discrete-factivity model for all predicates combined, with the mean log-odds for

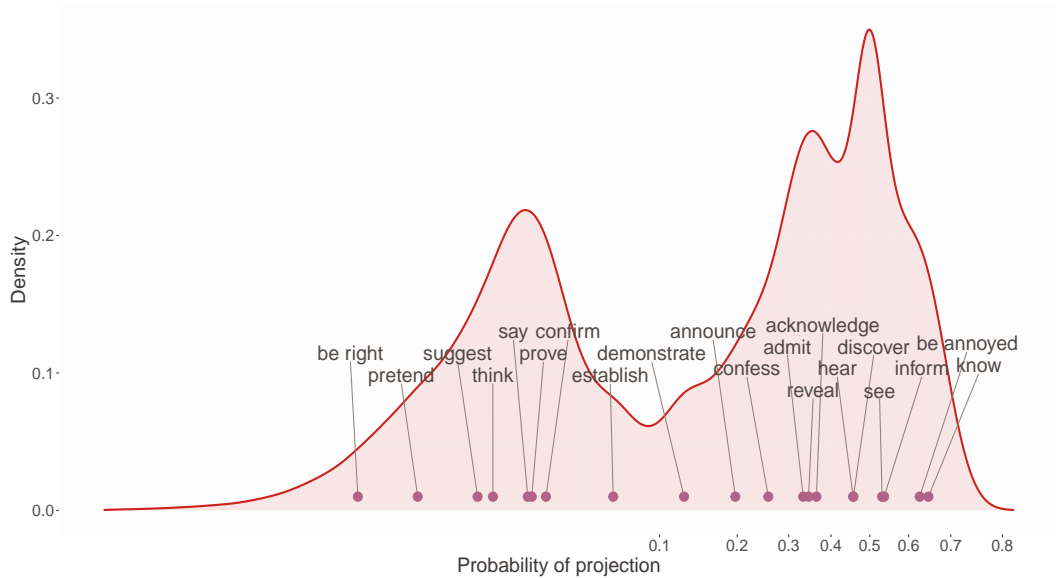


Figure 11: Density plot of the posterior probability of projection (with participant intercepts zeroed out) for the discrete-factivity model, for all predicates combined, scaled to log-odds space. Points represent the posterior mean log-odds associated with individual predicates (Spearman’s $r = 0.98^{***}$, when compared with the empirical means).

individual predicates represented toward the bottom. One can see that there are roughly two modes underneath which the means associated with individual predicates cluster. The left mode is around 0.018, or near zero. The right mode is close to a probability of around half. Indeed, one could make a cut between the predicates whose means appear to fall under the left versus the right mode. Such a cut would classify all of the predicates under Degen and Tonhauser’s “non-factive” category, except for *demonstrate*, as non-factive, along with *confirm* and *establish*; it would then classify the remaining predicates as optionally factive. Moreover, the contours visible in the right mode might suggest that more than one semantic class is active, insofar as such classes govern the frequency with which a given predicate triggers presupposition projection.

These observations are merely suggestive, however. We leave a detailed investigation of the semantic classes organizing the lexical knowledge of factivity for the future, noting merely that the extension of Grove and Bernardy’s framework proposed here provides a natural way to integrate uncertainty over predicate classes and their inferential effects into a model that connects the compositional semantics to experimental data in an unbroken chain.

7 Conclusion

As a whole, the results presented here can be taken to motivate a fairly traditional view of factivity, of the kind originally advocated by Kiparsky and Kiparsky (1970), Karttunen (1971), inter alia. Some predicates may be understood to trigger a presupposition that the clause they

select is true. The key departure from this tradition we would advocate, based on our results (and following Degen and Tonhauser), is in the particular classification of predicates which researchers ought to appeal to. Indeed, *none* of the predicates which Degen and Tonhauser investigate appear to be assigned a factive interpretation in all of their uses; rather, all seem to be associated with some degree of metalinguistic uncertainty about their status as factive. For many predicates, such as *think*, the degree of uncertainty is fairly trivial, fixing a near-zero probability of being factive. This is natural: if people are Bayesian reasoners about the knowledge they maintain about the world, including its linguistic conventions, some uncertainty about the semantic properties of linguistic expressions will be an essential feature of that knowledge.

References

- {Stan Development Team}. 2023. *Stan Modeling Language Users Guide and Reference Manual*, 2.32. Tech. rep.
- Abrusán, Márta. 2011. Predicting the presuppositions of soft triggers. *Linguistics and Philosophy* 34.6, pp. 491–535.
- Abrusán, Márta. 2016. Presupposition cancellation: explaining the ‘soft–hard’ trigger distinction. *Natural Language Semantics* 24.2, pp. 165–202. DOI: [10.1007/s11050-016-9122-7](https://doi.org/10.1007/s11050-016-9122-7).
- Abusch, Dorit. 2002. Lexical alternatives as a source of pragmatic presuppositions. *Semantics and Linguistic Theory*. Ed. by Brendan Jackson. Vol. 12. University of California, San Diego and San Diego State University, pp. 1–19.
- Abusch, Dorit. 2010. Presupposition triggering from alternatives. *Journal of Semantics* 27.1, pp. 37–80.
- Altmann, Gerry and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73.3, pp. 247–264.
- An, Hannah and Aaron White. 2020. The lexical and grammatical sources of neg-raising inferences. *Proceedings of the Society for Computation in Linguistics* 3.1, pp. 220–233. DOI: <https://doi.org/10.7275/yts0-q989>.
- Anand, Pranav and Valentine Hacquard. 2014. Factivity, Belief and Discourse. In *The Art and Craft of Semantics: A Festschrift for Irene Heim*. Ed. by Luka Crnić and Uli Sauerland. Vol. 1. MITWPL 70. MITWPL, pp. 69–90.
- Asudeh, Ash and Gianluca Giorgolo. 2020. *Enriched Meanings: Natural Language Semantics with Category Theory*. Oxford Studies in Semantics and Pragmatics. Oxford: Oxford University Press.
- Barker, Chris. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy* 25.1, pp. 1–36. DOI: [10.1023/A:1014346114955](https://doi.org/10.1023/A:1014346114955).

- Bergen, Leon, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9, ACCESS–ACCESS. DOI: [10.3765/sp.9.20](https://doi.org/10.3765/sp.9.20).
- Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, and Shalom Lappin. 2018. A Compositional Bayesian Semantics for Natural Language. *Proceedings of the First International Workshop on Language Cognition and Computational Models*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1–10.
- Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili. 2019a. Bayesian Inference Semantics: A Modelling System and A Test Suite. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 263–272. DOI: [10.18653/v1/S19-1029](https://doi.org/10.18653/v1/S19-1029).
- Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili. 2019b. Predicates as Boxes in Bayesian Semantics for Natural Language. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, pp. 333–337.
- Bernardy, Jean-Philippe, Rasmus Blanck, Stergios Chatzikyriakidis, and Aleksandre Maskharashvili. 2022. Bayesian Natural Language Semantics and Pragmatics. In *Probabilistic Approaches to Linguistic Theory*. Ed. by Jean-Philippe Bernardy et al. CSLI Publications.
- Charlow, Simon. 2014. On the semantics of exceptional scope. PhD Thesis. New York: New York University.
- Charlow, Simon. 2020. The scope of alternatives: indefiniteness and islands. *Linguistics and Philosophy* 43.4, pp. 427–472. DOI: [10.1007/s10988-019-09278-3](https://doi.org/10.1007/s10988-019-09278-3).
- Coppock, Elizabeth. 2018. Outlook-based semantics. *Linguistics and Philosophy* 41.2, pp. 125–164. DOI: [10.1007/s10988-017-9222-y](https://doi.org/10.1007/s10988-017-9222-y).
- De Marneffe, Marie-Catherine, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*. Vol. 23, pp. 107–124.
- Degen, Judith and Judith Tonhauser. 2021. Prior Beliefs Modulate Projection. *Open Mind* 5, pp. 59–70. DOI: [10.1162/opmi_a_00042](https://doi.org/10.1162/opmi_a_00042).
- Degen, Judith and Judith Tonhauser. 2022. Are there factive predicates? An empirical investigation. *Language* 98.3, pp. 552–591. DOI: [10.1353/lan.0.0271](https://doi.org/10.1353/lan.0.0271).
- Djäv, Kajsa and Hezekiah Akiva Bacovcin. 2017. Prosodic Effects on Factive Presupposition Projection. *Semantics and Linguistic Theory* 27.0, pp. 116–133. DOI: [10.3765/salt.v27i0.4134](https://doi.org/10.3765/salt.v27i0.4134).
- Djäv, Kajsa, Jérémy Zehr, and Florian Schwarz. 2018. Cognitive vs. emotive factives: An experimental differentiation. *Proceedings of Sinn und Bedeutung*. Vol. 21, pp. 367–386.
- Elliott, Patrick. 2020. Elements of Clausal Embedding. PhD thesis. University College London.

- Elliott, Patrick D. 2016. Explaining DPs vs. CPs without syntax. *Proceedings of the Fifty-first Annual Meeting of the Chicago Linguistic Society*. Ed. by Ksenia Ershova et al. Chicago: Chicago Linguistic Society, pp. 171–186.
- Elliott, Patrick D. 2022. A flexible scope theory of intensionality. *Linguistics and Philosophy*. DOI: [10.1007/s10988-022-09367-w](https://doi.org/10.1007/s10988-022-09367-w).
- Farudi, Annahita. 2007. An antisymmetric approach to Persian clausal complements. Ms., *University of Massachusetts, Amherst*.
- Frank, Michael C. and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336.6084, pp. 998–998. DOI: [10.1126/science.1218633](https://doi.org/10.1126/science.1218633).
- Gabry, Jonah and Rok Češnovar. 2023. *CmdStanR*. Tech. rep.
- Garnsey, Susan M. et al. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37.1, pp. 58–93.
- Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24.6, pp. 997–1016. DOI: [10.1007/s11222-013-9416-2](https://doi.org/10.1007/s11222-013-9416-2).
- Giannakidou, Anastasia. 1998. *Polarity sensitivity as (non) veridical dependency*. Vol. 23. John Benjamins Publishing.
- Giannakidou, Anastasia. 1999. Affective dependencies. *Linguistics and Philosophy* 22.4, pp. 367–421.
- Giannakidou, Anastasia. 2009. The dependency of the subjunctive revisited: Temporal semantics and polarity. *Lingua* 119.12, pp. 1883–1908.
- Giorgolo, Gianluca and Ash Asudeh. 2012. ⟨M, η, ★⟩ Monads for Conventional Implicatures. *Sinn und Bedeutung*. Ed. by Ana Aguilar Guevara, Anna Chernilovskaya, and Rick Nouwen. Vol. 16. MITWPL, pp. 265–278.
- Giorgolo, Gianluca and Ash Asudeh. 2014. One Semiring to Rule Them All. *CogSci 2014 Proceedings*.
- Givón, Talmy. 1973. The Time-Axis Phenomenon. *Language* 49.4, pp. 890–925.
- Goodman, Noah D. and Daniel Lassiter. 2015. Probabilistic Semantics and Pragmatics Uncertainty in Language and Thought. In *The Handbook of Contemporary Semantic Theory*. Ed. by Shalom Lappin and Chris Fox. John Wiley & Sons, Ltd, pp. 655–686. DOI: [10.1002/9781118882139.ch21](https://doi.org/10.1002/9781118882139.ch21).
- Goodman, Noah D. and Andreas Stuhlmüller. 2013. Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science* 5.1, pp. 173–184. DOI: <https://doi.org/10.1111/tops.12007>.

- Gordon, Peter and Jill Chafetz. 1990. Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition* 36.3, pp. 227–254.
- Grove, Julian. 2022. Presupposition projection as a scope phenomenon. *Semantics and Pragmatics* 15.15. DOI: [10.3765/sp.15.15](https://doi.org/10.3765/sp.15.15).
- Grove, Julian and Jean-Philippe Bernardy. 2023. Probabilistic Compositional Semantics, Purely. *New Frontiers in Artificial Intelligence*. Ed. by Katsutoshi Yada et al. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 242–256. DOI: [10.1007/978-3-031-36190-6_17](https://doi.org/10.1007/978-3-031-36190-6_17).
- Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9.3, pp. 183–221. DOI: [10.1093/jos/9.3.183](https://doi.org/10.1093/jos/9.3.183).
- Hooper, Joan B. 1975. On assertive predicates. In *Syntax and Semantics*. Ed. by John P. Kimball. Vol. 4. New York: Academy Press, pp. 91–124.
- Hooper, Joan B. and Sandra A. Thompson. 1973. On the Applicability of Root Transformations. *Linguistic Inquiry* 4.4, pp. 465–497.
- Jeong, Sunwoo. 2021. Prosodically-conditioned factive inferences in Korean: An experimental study. *Semantics and Linguistic Theory* 30.0, pp. 1–21. DOI: [10.3765/salt.v30i0.4798](https://doi.org/10.3765/salt.v30i0.4798).
- Kane, Benjamin, Will Gantt, and Aaron Steven White. 2022. Intensional Gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising. *Semantics and Linguistic Theory* 31.0, pp. 570–605. DOI: [10.3765/salt.v31i0.5137](https://doi.org/10.3765/salt.v31i0.5137).
- Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4.1, pp. 55–69. DOI: [10.1080/08351817109370248](https://doi.org/10.1080/08351817109370248).
- Karttunen, Lauri. 1974. Presuppositions and Linguistic Context. *Theoretical Linguistics* 1.1–3, pp. 181–194.
- Kastner, Itamar. 2015. Factivity mirrors interpretation: The selectional requirements of presuppositional verbs. *Lingua* 164, pp. 156–188.
- Kennedy, Christopher. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30.1, pp. 1–45. DOI: [10.1007/s10988-006-9008-0](https://doi.org/10.1007/s10988-006-9008-0).
- Kennedy, Christopher and Louise McNally. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language* 81.2, pp. 345–381. DOI: [10.1353/lan.2005.0071](https://doi.org/10.1353/lan.2005.0071).
- Kennedy, Christopher and Malte Willer. 2016. Subjective attitudes and counterstance contingency. *Semantics and Linguistic Theory* 26.0, pp. 913–933. DOI: [10.3765/salt.v26i0.3936](https://doi.org/10.3765/salt.v26i0.3936).
- Kennedy, Christopher and Malte Willer. 2022. Familiarity inferences, subjective attitudes and counterstance contingency: towards a pragmatic theory of subjective meaning. *Linguistics and Philosophy* 45.6, pp. 1395–1445. DOI: [10.1007/s10988-022-09358-x](https://doi.org/10.1007/s10988-022-09358-x).

- Kiparsky, Paul and Carol Kiparsky. 1970. FACT. In *Progress in Linguistics*. De Gruyter Mouton, pp. 143–173.
- Lassiter, Daniel. 2011. Vagueness as Probabilistic Linguistic Knowledge. *Vagueness in Communication*. Ed. by Rick Nouwen et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 127–150. DOI: [10.1007/978-3-642-18446-8_8](https://doi.org/10.1007/978-3-642-18446-8_8).
- Lassiter, Daniel and Noah D. Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* 194.10, pp. 3801–3836. DOI: [10.1007/s11229-015-0786-1](https://doi.org/10.1007/s11229-015-0786-1).
- Liang, Shen, Paul Hudak, and Mark Jones. 1995. Monad transformers and modular interpreters. *POPL '95 Proceedings of the 22nd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. New York, pp. 333–343.
- Liu, Fang and Evercita C Eugenio. 2018. A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research* 27.4, pp. 1024–1044. DOI: [10.1177/0962280216650699](https://doi.org/10.1177/0962280216650699).
- MacDonald, Maryellen C., Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101.4, p. 676.
- McBride, Conor and Ross Paterson. 2008. Applicative Programming with Effects. *Journal of Functional Programming* 18.1, pp. 1–13.
- Monroe, Will. 2018. Learning in the Rational Speech Acts model. PhD thesis. Stanford: Stanford University.
- Ozyildiz, Deniz. 2017. Attitude reports with and without true belief. *Semantics and Linguistic Theory*. Ed. by Dan Burgdorf et al. Vol. 27. Linguistic Society of America, pp. 397–417.
- Potts, Christopher et al. 2016. Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics* 33.4, pp. 755–802. DOI: [10.1093/jos/ffv012](https://doi.org/10.1093/jos/ffv012).
- Qing, Ciyang, Noah D. Goodman, and Daniel Lassiter. 2016. A rational speech-act model of projective content. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society: Recognising and representing events*. The Cognitive Science Society, pp. 1110–1115.
- Roberts, Craige. 2012. Information Structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5, 6:1–69. DOI: [10.3765/sp.5.6](https://doi.org/10.3765/sp.5.6).
- Ross, Alexis and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2230–2240. DOI: [10.18653/v1/D19-1228](https://doi.org/10.18653/v1/D19-1228).
- Roussou, Anna. 2010. Selecting complementizers. *Lingua* 120.3, pp. 582–603.
- Shan, Chung-chieh. 2002. Monads for natural language semantics. *arXiv:cs/0205026*.

- Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117.6, pp. 1034–1056. DOI: [10.1016/j.lingua.2006.05.006](https://doi.org/10.1016/j.lingua.2006.05.006).
- Simons, Mandy, David Beaver, et al. 2017. The Best Question: Explaining the Projection Behavior of Factives. *Discourse Processes* 54.3, pp. 187–206.
- Simons, Mandy, Judith Tonhauser, et al. 2010. What projects and why. *Semantics and Linguistic Theory*. Ed. by Nan Li and David Lutz. Vol. 20. University of British Columbia and Simon Fraser University: Linguistic Society of America, pp. 309–327. DOI: [10.3765/salt.v20i0.2584](https://doi.org/10.3765/salt.v20i0.2584).
- Tonhauser, Judith. 2016. Prosodic cues to presupposition projection. *Semantics and Linguistic Theory* 26.0, pp. 934–960. DOI: [10.3765/salt.v26i0.3788](https://doi.org/10.3765/salt.v26i0.3788).
- Tonhauser, Judith, David I. Beaver, and Judith Degen. 2018. How Projective is Projective Content? Gradience in Projectivity and At-issueness. *Journal of Semantics* 35.3, pp. 495–542. DOI: [10.1093/jos/ffy007](https://doi.org/10.1093/jos/ffy007).
- Trueswell, John C., Michael K. Tanenhaus, and Christopher Kello. 1993. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.3, p. 528.
- Unger, Christina. 2012. Dynamic Semantics as Monadic Computation. *New Frontiers in Artificial Intelligence*. Ed. by Manabu Okumura, Daisuke Bekki, and Ken Satoh. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 68–81. DOI: [10.1007/978-3-642-32090-3_7](https://doi.org/10.1007/978-3-642-32090-3_7).
- Varlokosta, Spyridoula. 1994. Issues in Modern Greek Sentential Complementation. PhD thesis. University of Maryland, College Park.
- Vehtari, Aki et al. 2023. *loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*.
- Von Fintel, Kai and Irene Heim. 2021. *Intensional semantics*. MIT.
- Watanabe, Sumio. 2013. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research* 14.27, pp. 867–897.
- White, Aaron S., Valentine Hacquard, and Jeffrey Lidz. 2018. Semantic Information and the Syntax of Propositional Attitude Verbs. *Cognitive Science* 42.2, pp. 416–456. DOI: [10.1111/cogs.12512](https://doi.org/10.1111/cogs.12512).
- White, Aaron Steven. 2019. Lexically triggered veridicality inferences. In *Handbook of Pragmatics*. Vol. 22. John Benjamins Publishing Company, pp. 115–148.
- White, Aaron Steven. 2021. On believing and hoping whether. *Semantics and Pragmatics* 14.6, pp. 1–18. DOI: [10.3765/sp.14.6](https://doi.org/10.3765/sp.14.6).
- White, Aaron Steven and Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory* 26.0, pp. 641–663. DOI: [10.3765/salt.v26i0.3819](https://doi.org/10.3765/salt.v26i0.3819).

- White, Aaron Steven and Kyle Rawlins. 2018a. Question agnosticism and change of state. *Proceedings of Sinn und Bedeutung* 21.2, pp. 1325–1342.
- White, Aaron Steven and Kyle Rawlins. 2018b. The role of veridicality and factivity in clause selection. *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.
- White, Aaron Steven, Rachel Rudinger, et al. 2018. Lexicosyntactic Inference in Neural Models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4717–4724. DOI: [10.18653/v1/D18-1501](https://doi.org/10.18653/v1/D18-1501).

A Full model specifications

Models were fit using Stan’s Hamiltonian Markov chain Monte Carlo sampling algorithm. For each model of Degen and Tonhauser’s data, as well as each model of either the bleached or templatic data, we obtained 6,000 posterior samples of the model parameters, following 6,000 burn-in samples, on four chains each. For each model of our replication experiment data, we obtained 24,000 posterior samples of the model parameters, following 24,000 burn-in samples, on four chains each.

A.1 The norming model

We characterize our model of the norming data as a probabilistic program, with the following structure, given data $\mathbf{y}_{\text{norming}} : r^{\frac{n_{\text{context}}}{2} * n_{\text{participant}}}$, where n_{context} and $n_{\text{participant}}$ are the number of contexts and participants, respectively, featured in the experiment. That is, each participant saw half of the available contexts, where each complement clause from the projection experiment was rated in conjunction with either a low-prior fact or a high-prior fact.

We encode the certainties for contexts as parameters ω on a log-odds scale, with participant random intercepts ϵ added to these parameters before they are mapped to transformed parameters w for certainty on the unit interval. Normal priors centered at zero are placed on the participant intercepts, as well as the log-odds parameters for contexts; the standard deviations (σ_{ϵ} and σ_{ω}) of these normal distributions are, in turn, given exponential hyperpriors. Finally, the likelihood for our model is given by a normal distribution centered at the certainty, whose standard deviation σ_e is parameterized with a prior uniform on the unit interval, truncated to the unit interval. We use $w_{i,j}$ to denote the parameter encoding the certainty for participant j , given context i .

We point out an important notational convention, which pertains to all of the model specifications we give here. We use the operator

$$D_{(\cdot)} : P\alpha \rightarrow \alpha \rightarrow r$$

to obtain a density (or mass, as the case may be) function on α ’s from a probabilistic program

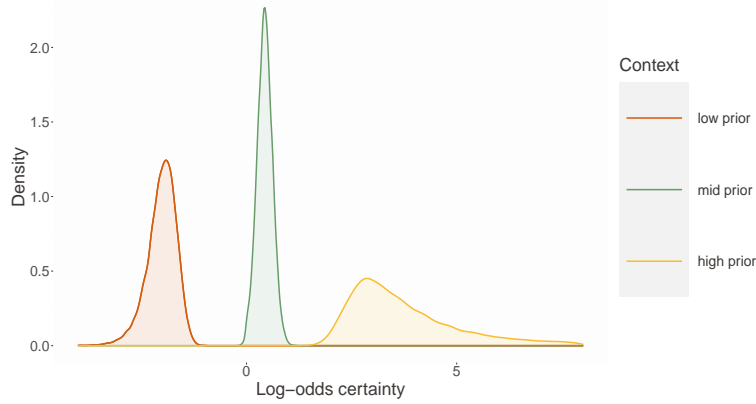


Figure 12: Density plots of the posterior log-odds certainty (with participant intercepts zeroed out) for three items in Degen and Tonhauser’s (2021) norming task. Low and high priors are for *Grace visited her sister*, given the facts *Grace hates her sister* and *Grace loves her sister*, respectively. Mid prior is for *Sophia got a tattoo*, given the fact *Sophia is a hipster*.

that returns α ’s as values. Thus if m returns, say, tuples of real numbers, then we may obtain the density (or mass) that m assigns to the tuple x as $D_m(x)$.

The model of the norming data can be presented compactly as follows:

$$\begin{aligned}
 \text{norming} &: P(r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{context}}} \times r^2) \\
 \text{norming} &= \sigma_{\omega} \sim \text{Exponential}(1) \\
 &\quad \sigma_{\epsilon} \sim \text{Exponential}(1) \\
 &\quad \sigma_e \sim \text{Uniform}(0, 1) \\
 &\quad \omega \sim \mathcal{N}(0, \sigma_{\omega}) \\
 &\quad \epsilon \sim \mathcal{N}(0, \sigma_{\epsilon}) \\
 &\quad \text{factor}(D_{\mathcal{N}(w, \sigma_e) \uparrow [0,1]}(\mathbf{y}_{\text{norming}})) \\
 &\quad \langle \omega, \epsilon, \sigma_{\omega}, \sigma_{\epsilon}, \sigma_e \rangle \\
 &\quad \text{where } w_{i,j} = \text{logit}^{-1}(\omega_i + \epsilon_j)
 \end{aligned}$$

The parameters ω encode a log-odds certainty rating for each item. We obtain prior distributions for these parameters in our models of factivity by extracting their marginal posterior distributions from our norming model and, for each item (i.e., each parameter of ω), taking a normal distribution with mean and variance equal to that of the posterior distribution. Density plots for three items are given in Figure 12 (see Figure 14 of Appendix B for all items).

A.2 The factivity models

We now provide our four models of factivity and prior world knowledge, which we fit to Degen and Tonhauser’s projection experiment data. In specifying each one, we use μ_{ω} and σ_{ω} to denote the means and standard deviations, respectively, of the marginal posterior distributions of the log-odds certainty ratings for the contexts assessed in the norming experiment.

In each model specification, $\mathbf{y}_{\text{projection}} : r^{n_{\text{verb}}} * n_{\text{participant}}$ encodes the experimental data, since each participant saw each verb exactly once.

For each model, we encode the log-odds of projection for verbs, along with the log-odds certainties for contexts, as parameters \mathbf{v} and $\boldsymbol{\omega}$. Participant random intercepts $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are added to these parameters, respectively, before they are mapped to transformed parameters \mathbf{v} and $\boldsymbol{\omega}$ on the unit interval. Normal priors centered at zero are placed on the participant intercepts, as well as the log-odds parameters for verbs; the standard deviations ($\sigma_{\boldsymbol{\epsilon}_1}$, $\sigma_{\boldsymbol{\epsilon}_2}$, and $\sigma_{\mathbf{v}}$) of these normals are, in turn, given exponential hyper-priors. Finally, the likelihoods for our models are again given by normal distributions truncated to the unit interval, and whose standard deviation σ_e is parameterized by a prior uniform on the unit interval. The mean $\boldsymbol{\theta}$ of this truncated normal likelihood varies by model, as we show next. In general, we use $\mathbf{v}_{i,j,k}$ and $\boldsymbol{\omega}_{i,j,k}$ to denote the parameters encoding the probability of projection and certainty, respectively, for participant k , given verb i and context j .

A.2.1 The discrete-factivity model

The discrete-factivity model defines the parameters $\boldsymbol{\theta}$ as either 1 or the certainty determined by world knowledge, depending on whether or not the relevant predicate's complement clause projects. This definition of $\boldsymbol{\theta}$ is justified by the following fact, given some fixed τ_1 :

Fact 1.

$$\mathbb{P} \left(\begin{array}{c} \tau_2 \sim \text{Bernoulli}(p) \\ \tau_1 \vee \tau_2 \end{array} \right) = \mathbb{1}(\tau_1) + \mathbb{1}(\neg\tau_1) * p$$

In other words, a given predicate's complement projects *or* it doesn't project; if it doesn't, then the prior certainty determined by world knowledge takes the reins. This yields the following model specification:

$$\text{discrete-factivity} : \mathbb{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^3)$$

$$\text{discrete-factivity} = \sigma_{\mathbf{v}} \sim \text{Exponential}(1)$$

$$\sigma_{\boldsymbol{\epsilon}_1} \sim \text{Exponential}(1)$$

$$\sigma_{\boldsymbol{\epsilon}_2} \sim \text{Exponential}(1)$$

$$\sigma_e \sim \text{Uniform}(0, 1)$$

$$\mathbf{v} \sim \mathcal{N}(0, \sigma_{\mathbf{v}})$$

$$\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \sigma_{\boldsymbol{\omega}})$$

$$\boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_1})$$

$$\boldsymbol{\epsilon}_2 \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_2})$$

$$\boldsymbol{\tau}_{\mathbf{v}} \sim \text{Bernoulli}(\mathbf{v})$$

$$\text{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \sigma_e) \top [0,1]}(\mathbf{y}_{\text{projection}}))$$

$$\langle \mathbf{v}, \boldsymbol{\omega}, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \sigma_{\mathbf{v}}, \sigma_{\boldsymbol{\epsilon}_1}, \sigma_{\boldsymbol{\epsilon}_2}, \sigma_e \rangle$$

$$\text{where } \mathbf{v}_{i,j,k} = \text{logit}^{-1}(\mathbf{v}_i + \boldsymbol{\epsilon}_{1k})$$

$$\boldsymbol{\omega}_{i,j,k} = \text{logit}^{-1}(\boldsymbol{\omega}_j + \boldsymbol{\epsilon}_{2k})$$

$$\boldsymbol{\theta}_{i,j,k} = \mathbb{1}(\boldsymbol{\tau}_{\mathbf{v}_{i,j,k}}) + \mathbb{1}(\neg\boldsymbol{\tau}_{\mathbf{v}_{i,j,k}}) * \boldsymbol{\omega}_{i,j,k}$$

A.2.2 The wholly-gradient model

The wholly-gradient model sets each parameter $\theta_{i,j,k}$ equal to $v_{i,j,k} + (1 - v_{i,j,k}) * w_{i,j,k}$, an encoding justified by the following fact:

Fact 2.

$$\mathbb{P} \left(\begin{array}{l} \tau_1 \sim \text{Bernoulli}(p) \\ \tau_2 \sim \text{Bernoulli}(q) \\ \tau_1 \vee \tau_2 \end{array} \right) = p + (1 - p) * q$$

Under this model, presupposition projection is genuinely gradient, since it adds directly to the certainty that the relevant complement clause is true, giving it a *boost* (albeit not all the way to 1).

$$\text{wholly-gradient} : P(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^3)$$

$$\begin{aligned} \text{wholly-gradient} = & \sigma_v \sim \text{Exponential}(1) \\ & \sigma_{\epsilon_1} \sim \text{Exponential}(1) \\ & \sigma_{\epsilon_2} \sim \text{Exponential}(1) \\ & \sigma_e \sim \text{Uniform}(0, 1) \\ & v \sim \mathcal{N}(0, \sigma_v) \\ & \omega \sim \mathcal{N}(\mu_\omega, \sigma_\omega) \\ & \epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}) \\ & \epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}) \\ & \text{factor}(D_{\mathcal{N}(\theta, \sigma_e) \top [0,1]}(\mathbf{y}_{\text{projection}})) \\ & \langle v, \omega, \epsilon_1, \epsilon_2, \sigma_v, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_e \rangle \end{aligned}$$

$$\begin{aligned} \text{where } v_{i,j,k} &= \text{logit}^{-1}(v_i + \epsilon_{1k}) \\ w_{i,j,k} &= \text{logit}^{-1}(\omega_j + \epsilon_{2k}) \\ \theta_{i,j,k} &= v_{i,j,k} + (1 - v_{i,j,k}) * w_{i,j,k} \end{aligned}$$

A.2.3 The discrete-world model

The discrete-world model is defined similarly to the discrete-factivity model, except by alternating which parameters are taken to make discrete versus gradient contributions to the response. Now, world knowledge affects the certainty discretely, producing values of either 0 or 1. Meanwhile, if the certainty is 0, the factivity of the relevant predicate makes a gradient

contribution to the response.

$$\begin{aligned}
&\text{discrete-world} : P(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^3) \\
&\text{discrete-world} = \sigma_v \sim \text{Exponential}(1) \\
&\quad \sigma_{\epsilon_1} \sim \text{Exponential}(1) \\
&\quad \sigma_{\epsilon_2} \sim \text{Exponential}(1) \\
&\quad \sigma_e \sim \text{Uniform}(0, 1) \\
&\quad v \sim \mathcal{N}(0, \sigma_v) \\
&\quad \omega \sim \mathcal{N}(\mu_\omega, \sigma_\omega) \\
&\quad \epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}) \\
&\quad \epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}) \\
&\quad \tau_w \sim \text{Bernoulli}(w) \\
&\quad \text{factor}(D_{\mathcal{N}(\theta, \sigma_e) \uparrow [0,1]}(\mathbf{y}_{\text{projection}})) \\
&\quad \langle v, \omega, \epsilon_1, \epsilon_2, \sigma_v, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_e \rangle \\
&\quad \text{where } v_{i,j,k} = \text{logit}^{-1}(v_i + \epsilon_{1k}) \\
&\quad \quad w_{i,j,k} = \text{logit}^{-1}(\omega_j + \epsilon_{2k}) \\
&\quad \quad \theta_{i,j,k} = \mathbb{1}(\tau_{w_{i,j,k}}) + \mathbb{1}(\neg \tau_{w_{i,j,k}}) * v_{i,j,k}
\end{aligned}$$

A.2.4 The wholly-discrete model

Finally, the wholly-discrete model generates parameters θ which are either 0 or 1, depending on two Bernoullis parameterized by the probabilities of projection and world-knowledge-derived certainties, respectively. Each parameter $\theta_{i,j,k}$ is thus 0 with probability $p = (1 - v_{i,j,k}) * (1 - w_{i,j,k})$, and 1 with probability $1 - p$.

$$\begin{aligned}
&\text{wholly-discrete} : P(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^3) \\
&\text{wholly-discrete} = \sigma_v \sim \text{Exponential}(1) \\
&\quad \sigma_{\epsilon_1} \sim \text{Exponential}(1) \\
&\quad \sigma_{\epsilon_2} \sim \text{Exponential}(1) \\
&\quad \sigma_e \sim \text{Uniform}(0, 1) \\
&\quad v \sim \mathcal{N}(0, \sigma_v) \\
&\quad \omega \sim \mathcal{N}(\mu_\omega, \sigma_\omega) \\
&\quad \epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}) \\
&\quad \epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}) \\
&\quad \tau_v \sim \text{Bernoulli}(v) \\
&\quad \tau_w \sim \text{Bernoulli}(w) \\
&\quad \text{factor}(D_{\mathcal{N}(\theta, \sigma_e) \uparrow [0,1]}(\mathbf{y}_{\text{projection}})) \\
&\quad \langle v, \omega, \epsilon_1, \epsilon_2, \sigma_v, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_e \rangle \\
&\quad \text{where } v_{i,j,k} = \text{logit}^{-1}(v_i + \epsilon_{1k}) \\
&\quad \quad w_{i,j,k} = \text{logit}^{-1}(\omega_j + \epsilon_{2k}) \\
&\quad \quad \theta_{i,j,k} = \mathbb{1}(\tau_{v_{i,j,k}} \vee \tau_{w_{i,j,k}})
\end{aligned}$$

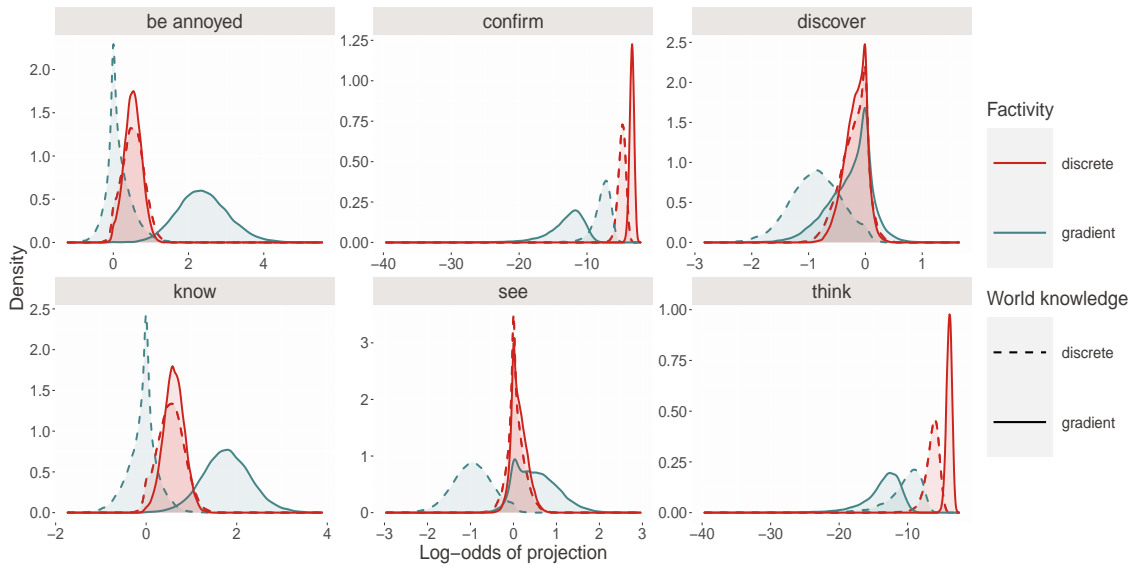


Figure 13: Density plots of the posterior log-odds of projection (with participant intercepts zeroed out) for all four models for six predicates from Degen and Tonhauser’s (2021) projection experiment.

A.3 The contentful evaluation model

To evaluate the four models using this data, we obtained, from each model, means μ_v and standard deviations σ_v of the marginal posterior log-odds of projection distributions for predicates, as well as means μ_ω and standard deviations σ_ω of the marginal posterior log-odds certainty distributions for contexts. We then used normal distributions with these means and standard deviations as prior distributions for the corresponding parameters in the models constructed for the evaluations. (See Figure 13 for density plots of these posterior distributions for six predicates, and Figure 15 of Appendix B for density plots of the posterior distributions for all predicates.)

Each evaluation has the following structure:

$$\begin{aligned}
 \text{replication-evaluation} &: P(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^3) \\
 \text{replication-evaluation} &= \sigma_{\epsilon_1} \sim \text{Exponential}(1) \\
 &\quad \sigma_{\epsilon_2} \sim \text{Exponential}(1) \\
 &\quad \sigma_e \sim \text{Uniform}(0, 1) \\
 &\quad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v) \\
 &\quad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}_\omega, \boldsymbol{\sigma}_\omega) \\
 &\quad \epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}) \\
 &\quad \epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}) \\
 &\quad \vdots \\
 &\quad \text{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}_e) \top [0,1]}(\boldsymbol{y}_{\text{replication}})) \\
 &\quad \langle \boldsymbol{v}, \boldsymbol{\omega}, \epsilon_1, \epsilon_2, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_e \rangle \\
 &\quad \text{where } \boldsymbol{v}_{i,j,k} = \text{logit}^{-1}(\boldsymbol{v}_i + \epsilon_{1k}) \\
 &\quad \quad \boldsymbol{w}_{i,j,k} = \text{logit}^{-1}(\boldsymbol{\omega}_j + \epsilon_{2k}) \\
 &\quad \quad \boldsymbol{\theta}_{i,j,k} = \dots
 \end{aligned}$$

The ellipsis are used to represent the parts of any given evaluation that are model-specific. For example, the line above factor would be ' $\boldsymbol{\tau}_v \sim \text{Bernoulli}(\boldsymbol{v})$ ' for the evaluation of the discrete-factivity model, and the definition of $\boldsymbol{\theta}_{i,j,k}$ would be $\mathbb{1}(\boldsymbol{\tau}_{v_{i,j,k}}) + \mathbb{1}(\neg \boldsymbol{\tau}_{v_{i,j,k}}) * \boldsymbol{w}_{i,j,k}$.

A.4 The non-contentful evaluation models

To evaluate the four models using both the bleached and templatic data, we used the means $\boldsymbol{\mu}_v$ and standard deviations $\boldsymbol{\sigma}_v$ of the marginal posterior log-odds of projection that we used for the evaluations on the replication experiment data. As before, we use normal distributions with these means and standard deviations as prior distributions for the corresponding parameters in the models constructed for the evaluations. Then, in each evaluation, we inferred a distribution over the parameters σ_ω and ω that regulate the certainty associated with either the bleached or the templatic context.

In particular, both the bleached and the templatic evaluations have the following struc-

ture, where ellipses, as above, indicate the unique aspects of each of the four models:

$$\begin{aligned}
 \text{non-contentful-evaluation} &: P(r^{n_{\text{verb}}} \times r \times r^{n_{\text{participant}}} \times r^{n_{\text{participant}}} \times r^3) \\
 \text{non-contentful-evaluation} &= \sigma_{\omega} \sim \text{Exponential}(1) \\
 &\quad \sigma_{\epsilon_1} \sim \text{Exponential}(1) \\
 &\quad \sigma_{\epsilon_2} \sim \text{Exponential}(1) \\
 &\quad \sigma_e \sim \text{Uniform}(0, 1) \\
 &\quad \mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v) \\
 &\quad \omega \sim \mathcal{N}(0, \sigma_{\omega}) \\
 &\quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}) \\
 &\quad \boldsymbol{\epsilon}_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}) \\
 &\quad \vdots \\
 &\quad \text{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}_e) \top [0,1]}(\mathbf{y}_{\text{non-contentful}})) \\
 &\quad \langle \mathbf{v}, \omega, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \sigma_{\omega}, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_e \rangle \\
 &\quad \text{where } \mathbf{v}_{i,j} = \text{logit}^{-1}(\mathbf{v}_i + \boldsymbol{\epsilon}_{1j}) \\
 &\quad \quad \mathbf{w}_{i,j} = \text{logit}^{-1}(\omega + \boldsymbol{\epsilon}_{2j}) \\
 &\quad \quad \boldsymbol{\theta}_{i,j} = \dots
 \end{aligned}$$

The data tuple $\mathbf{y}_{\text{non-contentful}}$ should be understood as either $\mathbf{y}_{\text{bleached}}$ or $\mathbf{y}_{\text{templatic}}$, depending on the evaluation performed.

B Plots

B.1 Posterior parameter distributions

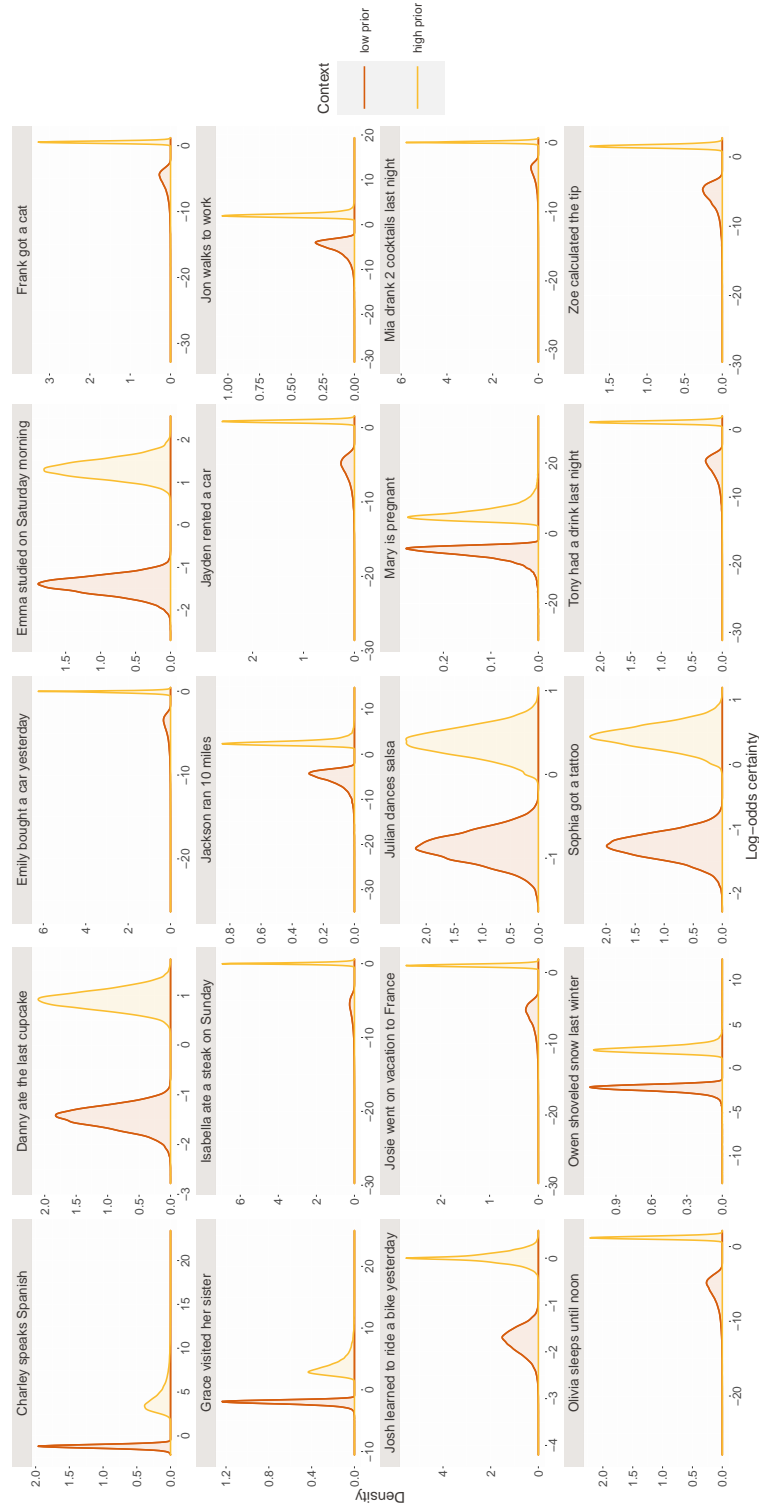


Figure 14: Density plots of the posterior log-odds certainty (with participant intercepts zeroed out) for all items in Degen and Tonhauser's (2021) norming task.

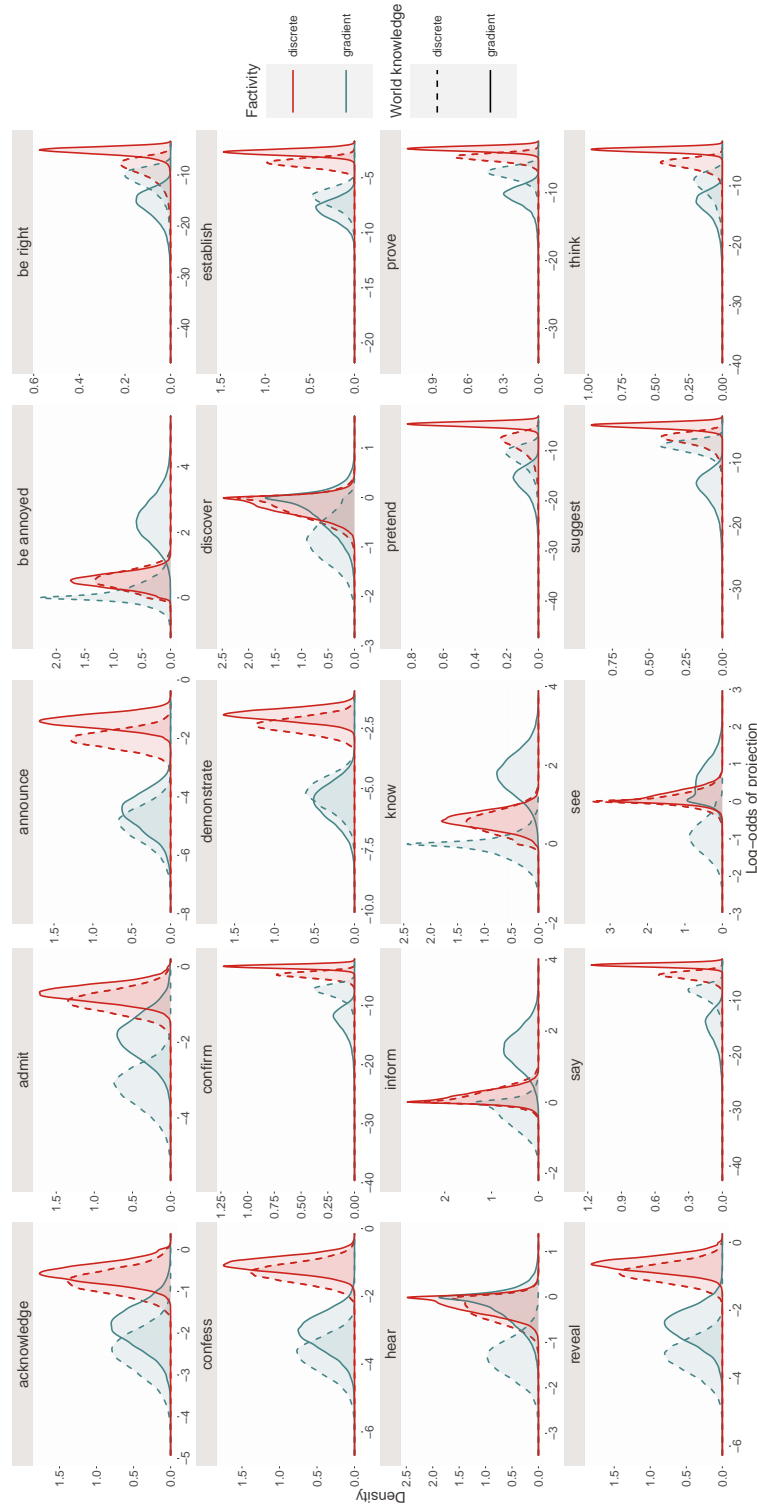


Figure 15: Density plots of the posterior log-odds of projection (with participant intercepts zeroed out) for all four models for all predicates in Degen and Tonhauser's (2021) projection experiment.

B.2 Posterior predictive distributions

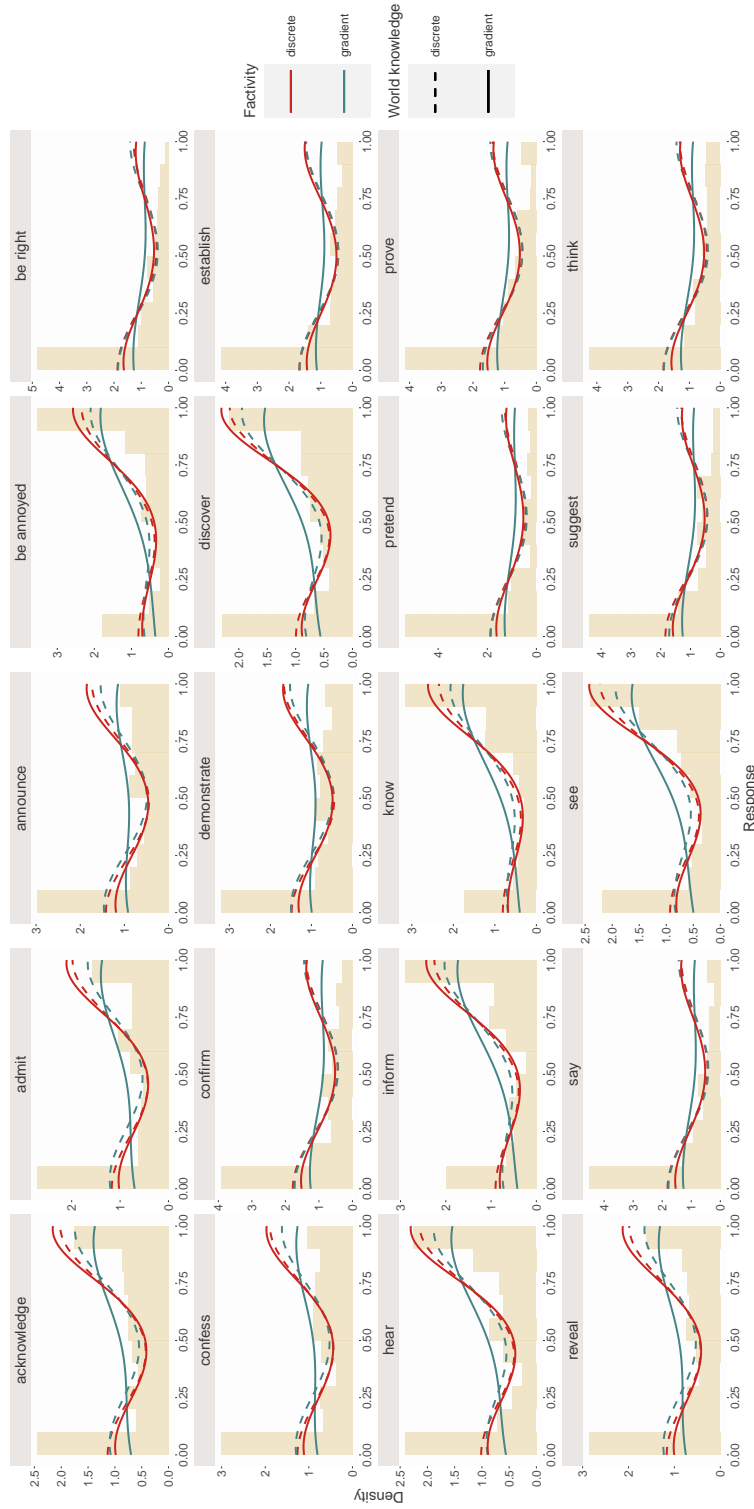


Figure 16: Posterior predictive distributions (with simulated participant intercepts) of all four models for all predicates in Degen and Tonhauser’s (2021) projection experiment, for all contexts combined. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

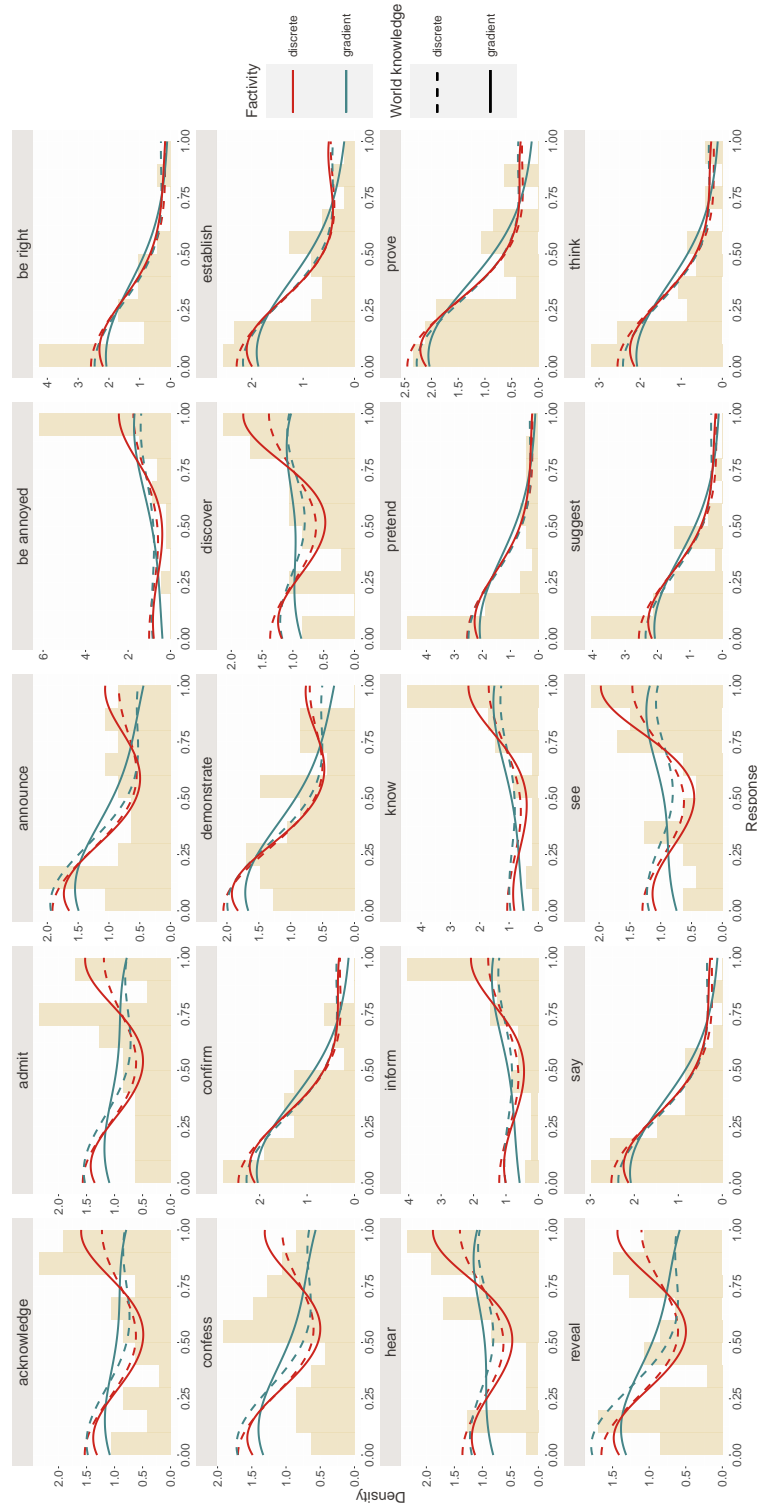


Figure 17: Posterior predictive distributions (with simulated participant intercepts) of all four model evaluations for all predicates in (3). Complement clause: *a particular thing happened*. Empirical distributions are represented by density histograms.

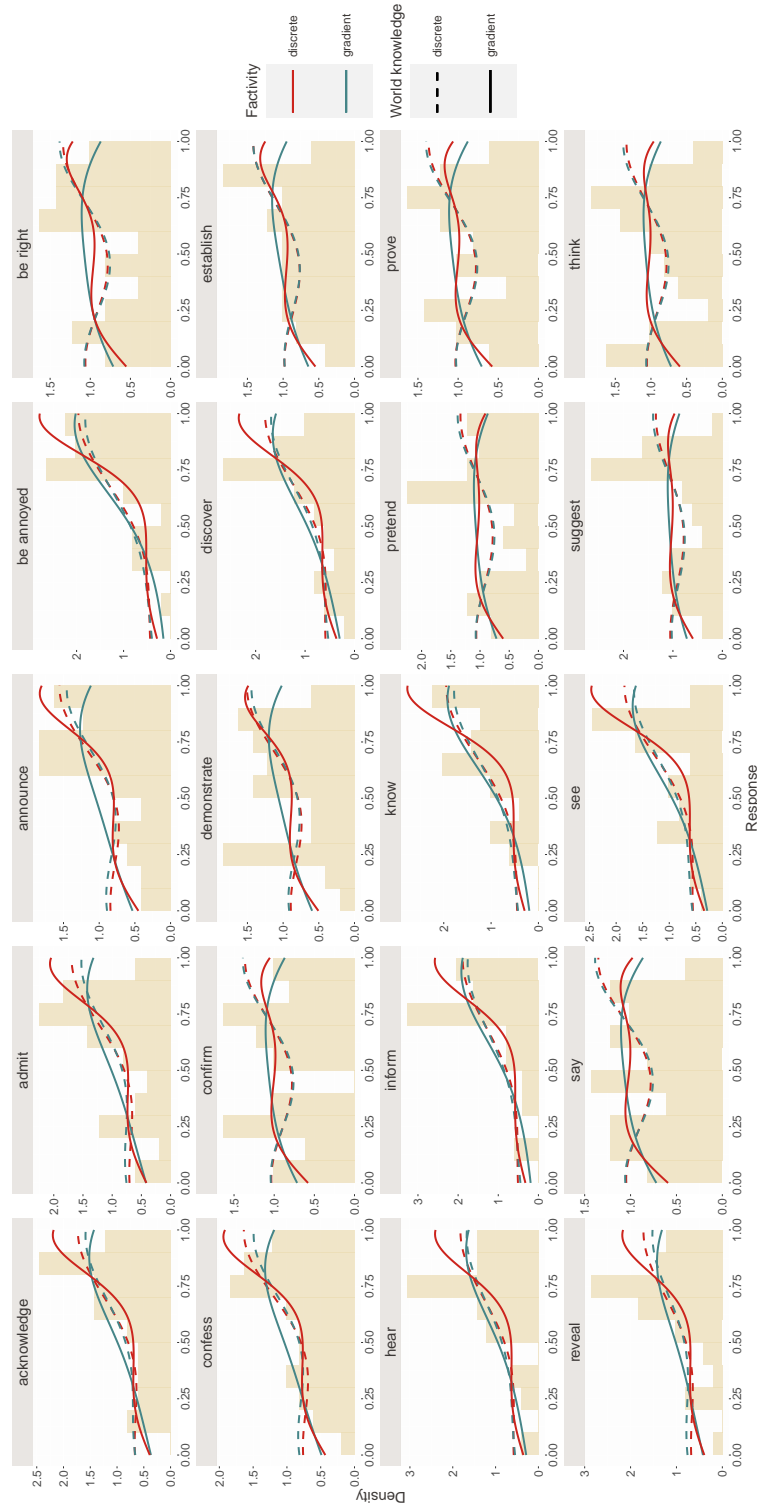


Figure 18: Posterior predictive distributions (with simulated participant intercepts) of all four model evaluations for all predicates in (3). Complement clause: *X happened*. Empirical distributions are represented by density histograms.