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' fits 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 implementing 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 the three datasets, we again find that models implementing the discreteness hypothesis fit the data better than models implementing the gradience hypothesis. We argue that these results favor an account on which factive presuppositions are fundamentally discrete in nature, and we discuss how this discreteness might be cashed out within both classical semantic accounts of factive predicates as well as accounts of presupposition projection that tie it to the question under discussion.

1 Introduction

Semantic theories aim to characterize the inferences that natural language expressions support and to account for at least a subset of those inferences: those that are necessary 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.

[*] We thank Judith Degen and Judith Tonhauser for making their experimental data and materials available. Thanks to Ash Asudeh, 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 2023 semantics seminar at the University of Chicago for feedback, as well as to attendees of the Linguistic Meaning Lab at Cornell.

Testing a semantic theory against inference judgment data thus requires auxiliary assumptions about the link between (some representation of) these factors and the theoretical 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 domain where large collections of inference judgments have become particularly important is presupposition projection and, in particular, factivity. A predicate is said to be *factive* if it is implicated in triggering *veridicality inferences* (i.e., inferences that its embedded clause is true) regardless of whether or not entailment canceling operators take scope over the 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 is known to be challenging, due to the influence a predicate's context of use exerts on the relevant veridicality inference (Karttunen 1971 *et seq.*). To better understand the factors driving factive inferences, it has become 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 concern 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 as far as to claim that the observed 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*, under which factivity is a discrete property or collection of properties. 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 of transparently relating the sorts of compositional analyses of expressions' meanings common in the formal semantics literature to probabilistic models characterizing distributions over inference judgments.

We formally define this framework in Section 3 before using it to carry out a comparison between the Fundamental Gradience Hypothesis and the Fundamental Discreteness Hypothesis in Sections 4 and 5. As we show, our framework allows us to precisely target where the two hypotheses make different predictions about the distribution of inference judgments across participants, making an apples-to-apples comparison feasible. Moreover, it allows us to do so using standard statistical model comparison metrics that balance out a model's fit to 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. One of these new datasets is a replication of the existing dataset, while the other two are novel. In Section 6, we argue that these results favor an account on which factive presuppositions are fundamentally discrete in nature, and we discuss how this discreteness might be cashed out within both classical semantic accounts of factive predicates as well as accounts that attempt to reduce presupposition projection to entirely pragmatic processes (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. Remarkably, the aggregate judgments of multiple speakers across such datasets show substantial gradience. This fact about distributions of inference judgments has garnered sustained focus.

In the domain of factivity, White and Rawlins (2018b) note gradience in the aggregate measures of different predicates' degrees of factivity, using data from the MegaVeridicality dataset (see Figure 1). In particular, they observe that there are no clear dividing lines among classes of predicates and suggest that speakers' gradient aggregate inferences about veridicality are likely influenced by the fine-grained semantics of particular verbs (*ibid*, p. 228; cp. Roberts and Simons to appear).[1]

In later work building on White and Rawlins 2018b, Degen and Tonhauser (2022) investigate the nature of inferential gradience in six experiments and argue that its persistence rebuts the hypothesis that there is a coherent class of factive predicates.

Our own modeling work uses data collected under the same experimental paradigm as that which Degen and Tonhauser employ, 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 may be responsible for the

[1] This gradience is not White and Rawlins's main focus: they are interested in the relationship between the veridicality inferences associated with particular predicates and those predicates' syntactic distributions—not those predicates' semantic classification—and, for their purposes, that relationship can be assessed without resolving what drives the gradience they observe in the aggregated inference judgments.

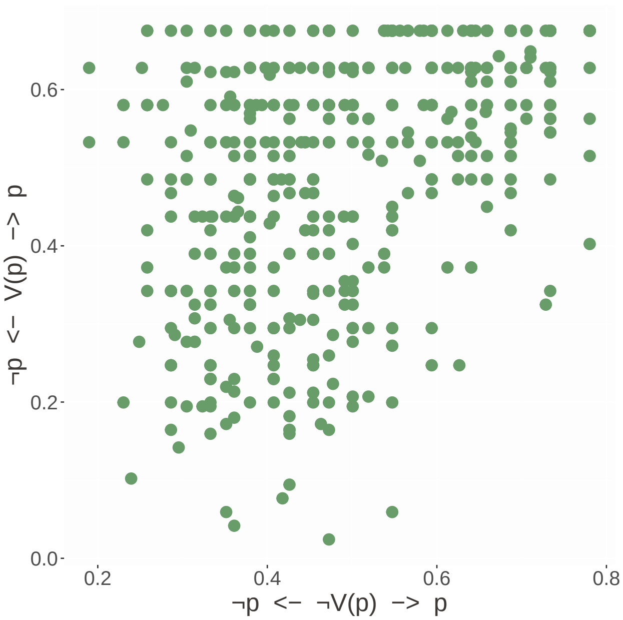


Figure 1: An aggregate measure of factivity, derived from the MegaVeridicality dataset of White and Rawlins 2018b. The y-axis corresponds to the mean ridit scores, across participants, of responses to prompts of the form *Someone (was) __ed that a particular thing happened. Did that thing happen?* with 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 green point is a predicate in either an active or a passive frame.

gradience observed in inference judgment tasks. We outline an array of explanatory choices that one can make in attempting to account for gradience, and we discuss some of the evidence that would be pertinent to these choices. In Section 2.4, we illustrate one particular kind of gradience seen in such tasks: that brought on by prior world knowledge. In Section 2.5, we present the two alternative hypotheses about gradience which constitute the main focus of this paper—that factive inferences are fundamentally discrete versus that they are fundamentally gradient—and we relate each hypothesis to the kinds of explanatory choices we introduce.

2.1 Measuring veridicality and factivity

In all of their experiments, Degen and Tonhauser (2022) focus on the set of twenty clause-embedding predicates listed in (3), which they group into hypothetical classes based on the prior literature on factivity (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 experimental data which they go on to collect, they suggest that insofar as the standard classificaton of predicates is correct, "…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). This prompt is meant 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, they give participants a variant of this task in which their answer is provided as a binary forced choice between *no* and *yes*, rather than a sliding scale response.

Degen and Tonhauser also claim that categories of predicates ought to emerge when one analyzes 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 follow this scenario with a prompt of the form in (7).

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

Depending on the experiment, participants either 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 give participants a scenario in which someone makes an utterance which should be contradictory if the relevant complement clause is entailed, as in (8).

(8) **Margeret**: "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.

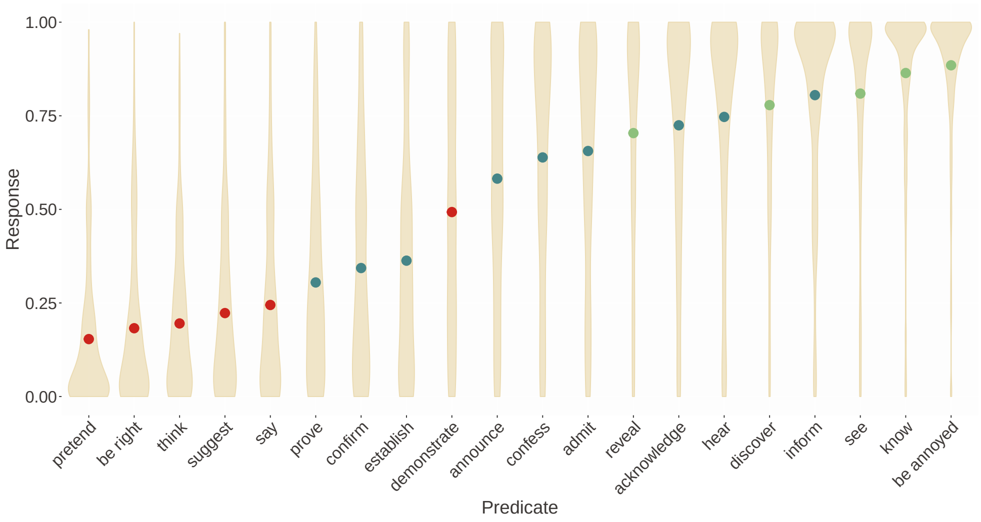


Figure 2: Verb means from Degen and Tonhauser's (2022) experiment 1a. "Non-factive" verbs are in red, "optionally factive" verbs are in teal, and "canonically factive" verbs are in green. Violin plots indicate the probability density of responses. (See Degen and Tonhauser 2022, Figure 2, p. 562.)

(9) "Is Margaret's utterance contradictory?"

Consistent with White and Rawlins's original observations, Degen and Tonhauser find that the patterns of inference across predicates in their experiments are gradient in nature for both projection and veridicality. Indeed, the degree to which predicates display projective inferences appears to evolve continuously from the least projective predicate (*pretend*) to the most projective predicate (*be annoyed*) when predicates are compared in terms of their mean ratings (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 with the veridical control items across all four experiments assessing veridicality.

2.2 Gradience in inference datasets

Degen and Tonhauser's results are consistent not only with White and Rawlins's original observations, but also with findings from adjacent domains. An and White (2020) observe similar gradience in neg-raising inferences, as captured in their MegaNegRaising dataset. Meanwhile, Kane, Gantt, and White (2022) note a similar pattern among the belief and desire inferences they obtain 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 the one hand, it is possible that "apparent gradience indicates that no formally represented lexical property controls whether a particular inference is triggered" (p. 572). On the other hand, "apparent gradience [may be] partly or wholly a product of the methods often used to collect inference judgments, and [it may be] that there are discrete, formally represented lexical properties that are active in triggering… inferences" (p. 572). To pursue this

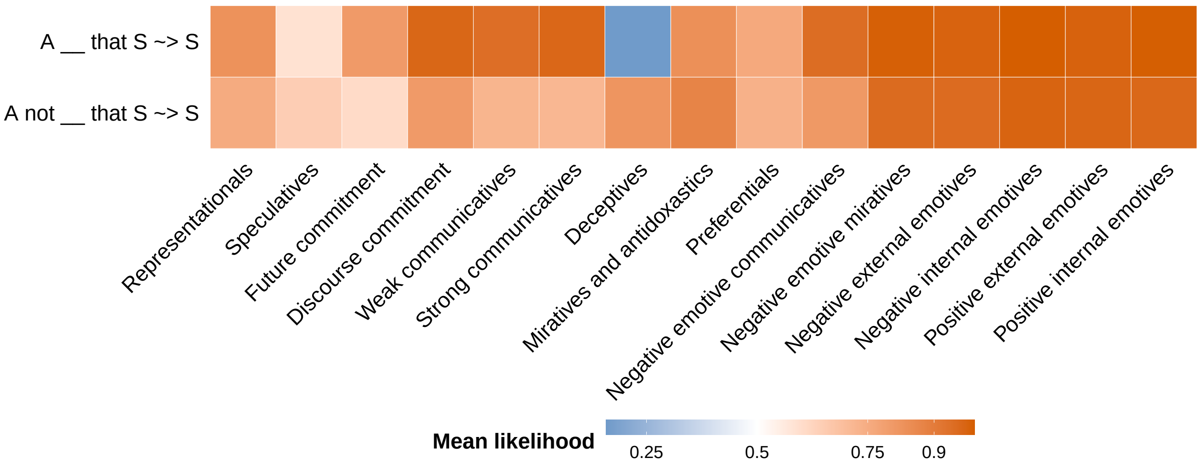


Figure 3: Veridicality inferences associated with each class found in Kane, Gantt, and White 2022. Kane, Gantt, and White provide labels for each class based on the predicates that occur in that class as well as the belief and desire inferences associated with that class. The top row and bottom row correspond 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.

question, they ask whether clear *patterns* emerge across the inference judgment datasets discussed above—MegaVeridicality, MegaNegRaising, and MegaIntensionality—by clustering predicates into classes sensitive 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 that correspond extremely closely to those that one would expect from prior work on clause-embedding predicates. As shown in Figure 3, these classes include a variety of factive subclasses that differ principally in the patterns of belief and desire inferences they are associated with. As one might expect from prior literature, the true factive subclasses tend to be emotive and include, 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 apparent gradience associated with certain classes, such as the non-emotive ones, as Degen and Tonhauser point out (their 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.

[2] The idea behind optimizing the predictability of predicates' syntactic distributions is that insofar as the classes to which a predicate belongs are predictive of the predicate's syntactic distribution, there is preliminary evidence that such classes are associated with distributionally active lexical representations.

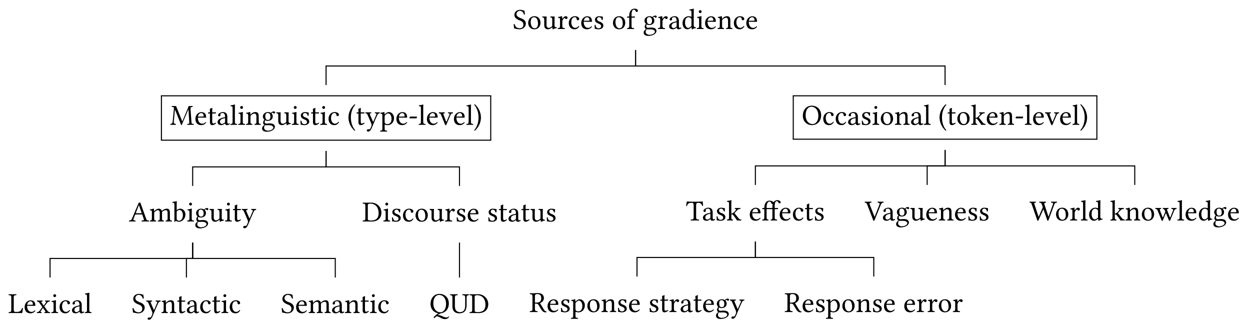


Figure 4: A partial taxonomy of the sources of gradience in inference judgment tasks.

2.3 Sources of gradience

How, then, should one account for gradience, whether it is due to knowledge about particular predicates or the classes those predicates fall into? To approach this question, we regiment the possible sources of gradience into two broad classes: *metalinguistic* (or *type-level*) sources and *occasional* (or *token-level*) sources.[3] A partial taxonomy of these sources is provided in Figure 4.[4]

2.3.1 Metalinguistic sources

Metalinguistic sources of gradience reflect uncertainty one might have about the *type* of speech act one is drawing an inference from. For example, in an utterance of the sentence *my uncle is running the race*, the verb *run* is ambiguous: the speaker's uncle might be a participant in the race, or he might be in charge of its logistics. If one is presented with an utterance of this sentence and is then asked, "On a scale of 1 to 10, how likely is it that my uncle has good managerial skills?", one might be tempted to pick a lower number—say, around 2—if *run* is interpreted in its locomotive sense, while one might be tempted to pick a higher number—say, around 8—if *run* is interpreted in its managerial sense. Indeed, if a hundred people are presented with an utterance of this sentence and are asked the same question, perhaps about half of them will interpret *run* in the locomotive sense, while the other half interpret it in the managerial sense. Thus, while the distribution of judgments will look bimodal—with one mode around 2 and the other mode around 8—the *average* judgment will be around 5. Gradience that stems from a semantic ambiguity is thus observable at the level of an entire experiment (a collection of inference judgments), rather than at the level of an individual trial (a single inference judgment). Likewise for the other categories of uncertainty classified as metalinguistic; for instance, uncertainty about discourse-level properties, such as the question under discussion, may lead to distinct inferences produced on different utterance occasions.

[3] The term 'occasional' is intended to contact Grice's (1989) distinction between *occasional* and *timeless* meaning.

[4] Indeed, this taxonomy is only partial. For instance, as Chris Kennedy (p.c.) points out, expressions displaying *open texture* seem likely to give rise to gradience (Waismann 1945), but it is potentially contentious whether open texture should be thought of as metalinguistic or occasional in nature.

2.3.2 Occasional sources

Occasional sources of gradience reflect uncertainty that persists about the status of an inference even after having fixed the expression being used to perform a particular speech act. One such kind of uncertainty comes from *vagueness*; for example, from an utterance of *my uncle is tall*, there can be uncertainty about whether or not my uncle is at least six feet tall. Another comes from *world knowledge*; given an utterance of *my uncle is running the race* (and having fixed *run* to its locomotive sense), there can still be uncertainty about how fast he runs, given that he is a runner. Crucially, such uncertainty would be reflected on *individual trials* of an inference judgment experiment. Moreover, metalinguistic gradience and occasional gradience can coexist: in the hypothetical inference judgment task described in the previous paragraph, there is metalinguistic uncertainty about the intended sense of the verb *run*, and there is occasional uncertainty about the inference that the referent of *my uncle* has good managerial skills, which itself stems from world knowledge. The latter uncertainty is reflected in the fact that the two modes (2 and 8) are slightly different from 0 and 10, respectively. Finally, occasional gradience in an experimental setting might also stem from task effects, such as one's physical response error or one's response strategy; for example, some experimental participants might hedge their responses so that they are never quite at the extremes of the relevant measurement scale.

2.3.3 The application to factivity

In the realm of factive inferences, one could adopt the view that gradience among individual predicates and classes of predicates is occasional in nature. This view would align with Tonhauser, Beaver, and Degen's *Gradient Projection Principle* (ex. 7, p. 499):

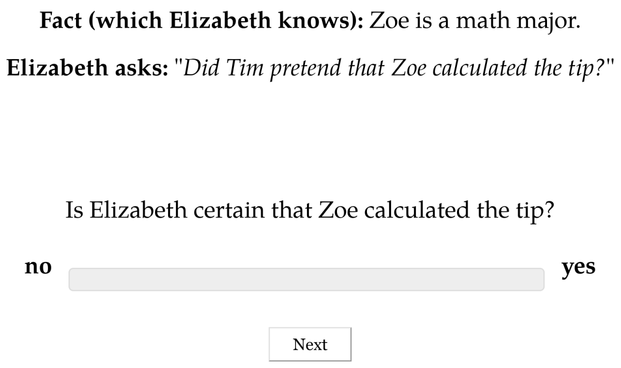
(10) **Gradient Projection Principle:** If content C is expressed by a constituent embedded under an entailment-canceling operator, then C projects to the extent that it is not at-issue.

On this view, (classes of) clause-embedding predicates license projective inferences which are themselves gradient.

Alternatively, one might take seriously the view that such gradience is actually metalinguistic. As Degen and Tonhauser put it, the latter view says 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 types of explanation so that we may quantitatively compare them.

2.4 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 projective inferences, focusing on the same twenty clause-embedding predicates as 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*:



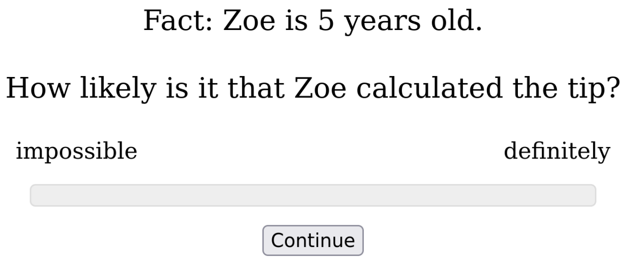
Fact (which Elizabeth knows): Zoe is a math major.

Elizabeth asks: "*Did Tim pretend that Zoe calculated the tip?*"

Is Elizabeth certain that Zoe calculated the tip?

The same twenty complement clauses as featured in 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 $r = 0.98$).

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 provided previously, but with the alternate low-probability fact:



Fact: Zoe is 5 years old.

How likely is it that Zoe calculated the tip?

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 which they obtain in their experiment investigating projective inferences.

Thus, at least one source of the 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?

2.5 Two hypotheses about gradience

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

(11) a. *The Fundamental Discreteness Hypothesis.* Factivity is a discrete 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 property distinguishing factive from non-factive occurrences of clause-embedding predicates. Rather, the gradient distinctions among predicates (and classes thereof) reflect different gradient contributions specific predicates make to inferences about the truth of their complement clauses.

According to the Fundamental Discreteness Hypothesis, the gradience among the predicates discussed above is metalinguistic (type-level) gradience. Individual occasions on which a predicate is used may be associated with uncertainty about whether or not the expression containing the predicate triggers a projective inference, but that uncertainty concerns which of the alternative interpretations of the expression should be selected. Alternative interpretations may be available for a variety of reasons: among other possibilities, (i) the predicate may have multiple senses—at least one that is implicated in triggering projection and at least one that is not; (ii) the predicate may occur in multiple structures—at least one that is implicated in triggering projection and at least one that is not;[5] or (iii) the predicate may be usable in contexts where the common ground—or some other construct, such as the question under discussion—entails the content of its embedded clause, as well as contexts where it does not (see Simons, Beaver, et al. 2017; Roberts and Simons to appear).[6] The gradience associated with particular classes of predicates that Kane, Gantt, and White observe might then indicate (i) a sort of regular polysemy among predicates within a class; (ii) that predicates within a class bias the resolution of syntactic ambiguity in similar ways; or (iii) that predicates within a class tend to show up in contexts with similarly structured common grounds.

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

[6] See Qing, Goodman, and Lassiter 2016 for an approach within the Rational Speech-Act framework (Frank and Goodman 2012; Goodman and Stuhlmüller 2013) along the lines of possibility (iii).

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).[7] That is, the gradience in the patterns of inference across predicates arises from an occasional (token-level) source. 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, under this hypothesis, there is no selection among alternative interpretations on particular occasions of use.

Importantly, both of these hypotheses concern what comprehenders do when they draw inferences from uses of a predicate. Does their behavior look more like ambiguity resolution or more like reasoning about vagueness or world knowledge? Importantly, this question is logically independent of the question of whether or not there is a semantic class of factive predicates (or subclasses thereof).

3 Probabilistic semantics

To state a theory of gradience precisely, it is useful to have a general method of integrating probabilistic reasoning into a compositional semantics of English. 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*.[8] Our reason for employing this framework is that it allows one to transparently relate the sorts of compositional analyses of expressions' meanings common in formal semantics to probabilistic models characterizing distributions over inference judgments.

Importantly, the framework allows one to precisely specify 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 which

[7] Tonhauser, Beaver, and Degen (2018) in fact consider two possibilities: (i) that "a listener's (or reader's) judgment that a content is projective to a certain extent means that the listener takes the speaker (or writer) to be committed to the content to that extent" (p. 498); and (ii) that "a listener's judgment that a content is projective to a certain extent reflects the probability with which they believe the speaker to be committed to the content" (pp. 498–499). The first of these, which they focus the majority of their discussion around, corresponds to our Fundamental Gradience Hypothesis. The second interpretation corresponds to our Fundamental Discreteness Hypothesis. For the remainder of the paper, we attribute the Fundamental Gradience Hypothesis to Tonhauser, Beaver, and Degen for a couple reasons: (a) we take that hypothesis to be the novel proposal they put forth; and (b) we reject the idea that possibility (ii) involves gradience in the linguistic representation in any interesting sense.

To sharpen the point, we note that one does not take the representation of *run* to be gradient simply because comprehenders presumably have (probably quite rich) knowledge about how frequently speakers intend *run* to have the motion sense versus the organizational sense (as potentially conditioned on a host of both linguistic and non-linguistic factors). This knowledge determines the inferences that comprehenders draw on the basis of the entailments of those senses; but we don't expect such inferences—like the linguistic representation itself—to be gradient in any interesting sense. We see no *a priori* reason to treat projective inferences any differently in this respect.

[8] See Giorgolo and Asudeh 2014; Asudeh and Giorgolo 2020 and Bernardy, Blanck, Chatzikyriakidis, and Lappin 2018 for related monadic approaches. The aforementioned works have somewhat different aims from Grove and Bernardy's, which are reflected in the distinct interfaces they supply.

both (a) precisely targets where they make different predictions about the distribution of inference judgments across experimental trials, and (b) 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 relevant background on Grove and Bernardy's framework before turning in Section 3.2 to our extension of it. Here we are able to finely delineate 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 also Bergen, Levy, and Goodman 2016; Potts et al. 2016; Monroe 2018 for a collection of related approaches). We illustrate our analysis of the distinction between these two forms of uncertainty by using an example involving vague adjectives; this allows us to highlight how Grove and Bernardy's framework can be used to approach occasional uncertainty. We then use the framework to give a minimalistic analysis of factivity in Section 3.3.

3.1 Denotations as probabilistic programs

The thrust of Grove and Bernardy 2023 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. This literature includes Shan's (2002) first introduction of monads to semanticists, 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 (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*. To model 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 model its denotation as contributing the entailment that the height of the individual x of which it is predicated is greater than some contextually determined height threshold d. Doing this, however, might result in a semantic representation like the following one, which involves an unbound degree variable (d):

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

There are different approaches to remedying this situation to be found in prior work. One approach assumes that the degree variable is existentially quantified—maybe, 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 will follow here—leaves the variable unbound and relies on the context to directly fix its value (see, e.g., Barker 2002; Kennedy 2007).[9] Among instances of the second kind of approach, many rely on probabilistic knowledge to constrain how the degree variable's value is fixed (Lassiter 2011; Goodman and Lassiter 2015; Lassiter and Goodman 2017; Bernardy, Blanck,

[9] Our 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 via contextual means.

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 provide such a probabilistic account. Their approach uses a monad to constrain the interpretation of the degree variable without tampering with the underlying compositional semantics of the adjective (and the rest of the sentence). Without worrying about its exact identity, we call this monad P here. Thus, P maps types, such as e, t, $e \rightarrow t$, $e \times t$, etc., onto types $\mathsf{P}e$, $\mathsf{P}t$, $\mathsf{P}(e \rightarrow t)$, $\mathsf{P}(e \times t)$, etc., which are inhabited by probabilistic programs. Because it is a monad, P comes with two monadic operators

$$\lambda m, k.\left(\begin{array}{l} x \sim m \\ k(x) \end{array}\right) \quad : \quad \mathsf{P}\alpha \rightarrow (\alpha \rightarrow \mathsf{P}\beta) \rightarrow \mathsf{P}\beta \qquad (\text{‘bind’})$$

$$\lambda x.\boxed{x} \quad : \quad \alpha \rightarrow \mathsf{P}\alpha \qquad (\text{‘return’})$$

which we describe in Sections 3.1.1 and 3.1.2.

First, a note of clarification about our use of the jargon ‘probabilistic program’. In our system, $\mathsf{P}\alpha$ is a type inhabited by λ-terms, just like all other types are. But whereas the meaning of a λ-term that inhabits a function type, $\alpha \rightarrow \beta$, is a function which, given something of type α, computes something of type β, the meaning of a λ-term that inhabits the type $\mathsf{P}\alpha$ is a computation of a random value of type α according to some probability distribution. How exactly this computation is carried out—whether, e.g., by analytically computing a closed-form representation of the density associated with an arbitrary value of type α, or by approximating this density by sampling values directly from the relevant distribution, or by yet some other method—is immaterial. The purely declarative language of λ-terms makes space to do one thing and one thing only: to *describe* the sort of computation of type $\mathsf{P}\alpha$ which must be carried out in order to compute a random value of type α. Describing such a computation answers questions such as, “What information is this random value conditional on?”, “Does it depend somehow on a value computed from a particular distribution?”, “Are there values it is blocked from taking on?”. That is, it provides a complete mathematical description of any random variable of interest in terms of the dependencies it enters into in virtue of the control flow of the program. By maintaining such indifference about implementation details, we are afforded a very useful abstraction that puts probabilistic programs on the same plane as other typed λ-terms. Moreover, it allows them to be integrated into the ambient calculus seamlessly, without changing its meaning. Meanwhile, the monadic interface endows probabilistic programs with the structure required to compose programs together, to register relevant dependencies, and to return values. Here is how.

3.1.1 Bind

The bind operator can be used to characterize the interpretation of parameters, like d above, by sequencing one probabilistic program with another that depends on a variable. Sequencing some program m of type $\mathsf{P}\alpha$ with a continuation of that program k of type $\alpha \rightarrow \mathsf{P}\beta$

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

can be understood as sampling a random value x of type α from m and then using x to construct the new probabilistic program $k(x)$ of type $\mathsf{P}\beta$.

The feature of bind making it indispensable is that it allows the values returned by one probabilistic program to feed into another probabilistic program. For example, let's say we want to describe a normal distribution whose mean and standard deviation are unknown. In other words, the parameters of such a normal distribution themselves take a distribution of some kind. This situation is easily described in terms of bind, which allows the following non-deterministically parameterized normal distribution to be encoded:

$$\begin{array}{l} \mu \sim \mu\mathtt{dist} \\ \sigma \sim \sigma\mathtt{dist} \\ \mathcal{N}(\mu, \sigma) \end{array}$$

Crucially, both μ and σ appear as random variables which are sampled from prior distributions, where they appear on the *left* side of a bind statement. Immediately underneath, they appear as parameters of the returned program $\mathcal{N}(\mu, \sigma)$. If a new value were to be bound to this returned program, then the parameters μ and σ would appear on the *right* side of the corresponding new bind statement, where they would determine the distribution being sampled from (see Section 3.1.4 for an example of this kind of situation). That a given parameter may occur on either side of a bind statement—e.g., that σ may appear on the left, bound by $\sigma\mathtt{dist}$, and then appear as a parameter of the new distribution $\mathcal{N}(\mu, \sigma)$ further down—is what makes the monadic interface so powerfully expressive. Complex interconnections may obtain, in which the random value computed by one program controls the probabilistic effect associated with another.

3.1.2 Return

The return operator allows ordinary logical meanings to be lifted into probabilistic programs associated with a *trivial* effect. It is trivial in the sense that it always computes the same value. For instance, sampling from $\boxed{\mathsf{j}} : \mathsf{P}e$ will always result in $\mathsf{j} : e$. In the parlance of probability theory, such programs describe *degenerate distributions*.

Return is useful in order to return a function of the values computed by some collection of (indexed) probabilistic programs. That is, given programs $m_1 : \mathsf{P}\alpha_1, m_2 : \alpha_1 \rightarrow \mathsf{P}\alpha_2, ..., m_n : \alpha_1 \times ... \times \alpha_{n-1} \rightarrow \mathsf{P}\alpha_n$, and some function $f : \alpha_1 \times ... \times \alpha_n \rightarrow \beta$, one can do

$$\begin{array}{l} x_1 \sim m_1 \\ x_2 \sim m_2(x_1) \\ \vdots \\ x_n \sim m_n(x_1, ..., x_{n-1}) \\ \boxed{f(x_1, ..., x_n)} \end{array}$$

to obtain the program of type $\mathsf{P}\beta$ which *applies* f to the values computed by the programs m_i.

3.1.3 The monad laws

Because P together with bind and return is a monad, it satisfies the laws in Figure 5. Among these laws, Left identity guarantees that transforming a value v via return creates a "pure" probabilistic program that just returns v; that is, it ensures that $\boxed{v}$ encodes a degenerate distribution. 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

Left identity *Right identity* *Associativity*

$$\begin{array}{l} x \sim \boxed{v} \\ k(x) \end{array} \quad = \quad k(v) \qquad\qquad \begin{array}{l} x \sim m \\ \boxed{x} \end{array} \quad = \quad m \qquad\qquad \begin{array}{l} y \sim \left(\begin{array}{l} x \sim m \\ n(x) \end{array} \right) \\ o(y) \end{array} \quad = \quad \begin{array}{l} x \sim m \\ y \sim n(x) \\ o(y) \end{array}$$

Figure 5: The monad laws

programs to be re-bracketed: if one samples y from a complex probabilistic program that contains a use of bind, one may also pull out the parts composing the program and instead sample y from the last part.

Together, the laws ensure a clean separation between the effectful, probabilistic aspects of a probabilistic program and the pure aspects of the program, i.e., those which only manipulate values. Indeed, the pure aspects are those pieces of a program which would normally be taken to be semantically relevant (see Charlow 2014; Charlow 2020 for extensive discussion)—what Charlow (2014) terms the "Fregean bread and butter" of compositional semantics. By maintaining a separation between a compositional semantic core and a probabilistic mantle, one may import existing semantic analyses into the probabilistic setting, where they become part of the core. At the same time, because monadic probabilistic programs are type-safe and lawful, one may reason about their behavior and, for example, derive equivalences to the Bayesian models which we present in Section 4.

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

To pump intuitions a bit, we look at the gradable adjective *tall*. To model the semantic values of adjectives like *tall*, Grove and Bernardy assume that $[\![tall]\!]$ is a probabilistic program of type $\mathsf{P}(e \to t)$.[10] 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:

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

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

Importantly, $\mathsf{thresholdPrior}$ can be anything, as long as it is of the right type, $\mathsf{P}r$. Its main use is to represent the constraints that the context—including comprehenders' prior beliefs—imposes on d. For instance, one could take the threshold to be distributed according to the underdetermined

[10] Since they take P to be the continuation monad with result type r (the type of real numbers), the semantic value of *tall* for them is of type $((e \to t) \to r) \to r$.

normal distribution discussed above:

$$\begin{aligned} & d \sim \left(\begin{array}{l} \mu \sim \mu\mathtt{dist} \\ \sigma \sim \sigma\mathtt{dist} \\ \mathcal{N}(\mu, \sigma) \end{array} \right) \\ & \boxed{\lambda x.\mathsf{height}(x) \geq d} \\ = \; & \mu \sim \mu\mathtt{dist} && \text{(by Associativity)} \\ & \sigma \sim \sigma\mathtt{dist} \\ & d \sim \mathcal{N}(\mu, \sigma) \\ & \boxed{\lambda x.\mathsf{height}(x) \geq d} \end{aligned}$$

Under this assumption, the height threshold is sampled from—i.e., *bound by*—the program that computes a normal distribution with unknown mean and standard deviation.

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, 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 $\mathsf{P}\alpha$ to values of type r (real numbers). To satisfy this requirement, they (at least implicitly) use an *expected value* operator:

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

Given a function f from values of type α to real numbers, we write $\mathbb{E}_{x \sim m}\left[f(x)\right]$ for the expected value of f, given the probability distribution over values of type α represented by m.[11] If m returns truth values—i.e., if it is of type $\mathsf{P}t$—it can be associated with a probability by taking the expected value of the indicator function $\mathbb{1} : t \rightarrow r$, which maps T ('true') to 1 and F ('false') to 0:

$$\begin{gathered} \mathbb{P} : \mathsf{P}t \rightarrow r \\ \mathbb{P}(m) = \mathbb{E}_{\tau \sim m}\left[\mathbb{1}(\tau)\right] \end{gathered}$$

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

$$\begin{aligned} [\![\textit{Jo is tall}]\!] &: \mathsf{P}t \\ [\![\textit{Jo is tall}]\!] &= \; d \sim \mathcal{N}(\mu, \sigma) \\ & \boxed{\mathsf{height}(\mathsf{j}) \geq d} \end{aligned}$$

we can use $\mathbb{P}$ to compute the probability

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

[11] An 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 determined by the probabilistic program m).

Thus, the probability that *Jo is tall* is true is equal to the probability that $\mathsf{height}(\mathsf{j}) \geq d$, where d is a normally distributed random variable with mean μ and standard deviation σ. Alternatively, it is the cummulative density function of the relevant distribution at the value corresponding to Jo's height.

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 (see von Fintel and Heim 2021 and references therein), or situations (Barwise and Perry 1983). In addition to providing parameters that determine the denotations of expressions, contexts provide values for contextual parameters; for example, 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*:

$$\begin{aligned}
[\![\mathit{tall}]\!] &: e \longrightarrow \kappa \longrightarrow t \\
[\![\mathit{tall}]\!] &= \lambda x, c.\mathsf{height}(c)(x) \geq \mathsf{d}_{\mathit{tall}}(c)
\end{aligned}$$

$\mathsf{height}(c)$ selects whichever component of c maps individuals to their heights, and $\mathsf{d}_{\mathit{tall}}(c)$ selects whichever component of c provides the contextual degree threshold relevant to determining the truth of the vague 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 may now be conveniently viewed as sets of contexts, or functions of type $\kappa \longrightarrow t$. Furthermore, following Grove and Bernardy (2023), the common ground may be viewed as a *distribution* over contexts, or a probabilistic program of type $\mathsf{P}\kappa$. To update the common ground with a proposition, Grove and Bernardy define a function `observe` in terms of a more primitive operation `factor`. The role of `factor` is to weight the distribution represented by the probabilistic program which follows it by some scalar value:[12]

$$\mathtt{factor} : r \longrightarrow \mathsf{P}\diamond$$

$$\mathtt{observe} : t \longrightarrow \mathsf{P}\diamond$$

$$\mathtt{observe}(\phi) \overset{\text{def}}{=} \mathtt{factor}(\mathbb{1}(\phi))$$

[12] In the continuation-based setting of Grove and Bernardy 2023, `factor` is defined as

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

so that it scales its continuation by the relevant factor. We maintain an abstract interface here so that our main points aren't obscured by implementation details.

$\diamond$ is the unit type; it is inhabited by a single value: the 0-tuple (also written '$\diamond$'). It therefore carries no interesting information. As a result, `factor` contributes only a probabilistic effect, rather than a value.

Given some common ground, $cg : \mathtt{P}\kappa$, one can update it with the proposition $\phi : \kappa \to t$ in terms of `observe`:

$$\begin{aligned} \mathtt{update} &: (\kappa \to t) \to \mathtt{P}\kappa \to \mathtt{P}\kappa \\ \mathtt{update}(\phi)(cg) &\overset{\text{def}}{=} \ c \sim cg \\ &\ \ \mathtt{observe}(\phi(c)) \\ &\ \ \boxed{c} \end{aligned}$$

Dynamizing propositions is thus a matter of *observing* them in the context provided by the relevant common ground. Note that our concept of a dynamic proposition is similar to the notion of a context-change potential (Heim 1982; Heim 1983). More generally, it is akin to update in a *distributive* dynamic semantics, in which contexts are updated pointwise; the main departure from such a system being that points are now weighted by probability densities.

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. We expound on this point now.

3.2 Our contribution: two levels of uncertainty

Our main contribution comes from 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 $\mathtt{P}\kappa$—we represent it as a probability distribution *over* probability distributions over contexts—i.e., as a program of type $\mathtt{P}(\mathtt{P}\kappa)$. By invoking the functor $\mathtt{P}$ twice, we are effectively providing two layers, or levels, of probabilistic uncertainty. Thus, not only is there a Montagovian core and a probabilistic mantle, but there is now an additional probabilistic crust.

We use the "inner" $\mathtt{P}$ to represent the uncertainty that is manifest on particular occasions of use and interpretation; that is, occasional (token-level) uncertainty. To recapitulate the discussion of Section 2, such uncertainty might arise because of linguistic expressions which are vague or subjective, or it may be uncertainty related to prior beliefs that people have about the world.

We use the "outer" $\mathtt{P}$ to represent metalinguistic (type-level) uncertainty. Although there may, in general, be uncertainty about the values of linguistic parameters that govern the meanings of expressions, by regulating this uncertainty on the outer layer, we take those values to be fixed on particular occasions of language use and interpretation. Thus, the outer $\mathtt{P}$ provides a distribution over possible *types* of occasions of use and interpretation—that is, which fix the values of parameters which are metalinguistically uncertain—while the inner $\mathtt{P}$ regulates residual uncertainty that arises on particular occasions of use and interpretation, even once the relevant type of occasion has been fixed.

Which phenomena should be tethered to which layer of uncertainty is, importantly, up for debate and should ultimately be settled empirically (though recall Figure 4 of Section 2 for a partial

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

Figure 6: The functor laws

conjecture in this direction). 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 reiterate 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) settled once the occasion of use is fixed, or (b) an inherent property of particular occasions of use and interpretation, 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 $(\cdot)^{\Updownarrow}$ ('map') allowing one to perform pure operations on the values returned by probabilistic programs, while keeping their probabilistic effects intact. $(\cdot)^{\Updownarrow}$ may be defined in terms of return and bind, as follows:

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

There are two laws defining functors, which are given in Figure 6.[13]

Crucially, functors are *composable*, meaning that we can take the composition of the functor P with itself to obtain the new functor $\mathsf{P} \circ \mathsf{P}$, whose map may be defined by composing the map of P with itself, i.e., $(\cdot)^{\Updownarrow\Updownarrow}$.[14] Old operations are easily recast in the current setting involving structured uncertainty, that is, by *mapping them* onto operations on higher-order probabilistic programs. Updates to the common ground, for instance, may be presented as follows:

$$\begin{aligned} \mathtt{update}_2 &: (\kappa \to t) \to \mathsf{P}(\mathsf{P}\kappa) \to \mathsf{P}(\mathsf{P}\kappa) \\ \mathtt{update}_2(\phi) &= \mathtt{update}(\phi)^{\Updownarrow} \end{aligned}$$

The second property of note is that, because $\mathsf{P} \circ \mathsf{P}$ 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 occasional uncertainty. As a result, it is possible for occasional uncertainty to remain even after questions of metalinguistic uncertainty have been settled; e.g., whether a semantically ambiguous expression has one interpretation or another. But by necessity, settling token-level uncertainty requires type-level uncertainty to already be settled.

[13] Note that either may be proved from the monad laws of Figure 5.

[14] Indeed, 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

$$(\circledast) : \mathsf{P}(\alpha \to \beta) \to \mathsf{P}\alpha \to \mathsf{P}\beta$$

called 'sequential application', which can apply an effectful function to an effectul 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 currently require.

We construct this asymmetry to model the general behavior of the two sources of uncertainty under investigation. For illustration, 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:

$$\llbracket \textit{Jo is \%\&-all} \rrbracket : \mathsf{P}(\mathsf{P}t)$$

$$\llbracket \textit{Jo is \%\&-all} \rrbracket = \; \tau \sim \texttt{Bernoulli}(0.7)$$

$$\begin{cases} \boxed{\begin{array}{l} d \sim \mathcal{N}(\mu_t, \sigma_t) \\ \boxed{\mathsf{height(j)} \geq d} \end{array}} & \tau \\ \boxed{\begin{array}{l} d \sim \mathcal{N}(\mu_s, \sigma_s) \\ \boxed{\mathsf{size(j)} \leq d} \end{array}} & \neg\tau \end{cases}$$

According to this interpretation, the meaning of *Jo is %&-all* depends on the Bernoulli-distributed variable $\tau : t$. If τ is T (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 F (which occurs with a probability of 0.3), then the interpretation 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 occasional uncertainty, encoded by a normal distribution over degrees of height or size, respectively. Thus, the probabilistic effects encoding occasional uncertainty depend on those encoding metalinguistic uncertainty (about the value of τ, in particular). But the former effects cannot, in turn, influence the latter effects, 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 a characterization of factivity within this two-tiered probabilistic setting.

3.3 The meaning of factivity

In general, we assume that clause-selecting predicates entail the complement clauses they select with some probability.[15] 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:

$$\llbracket \textit{know} \rrbracket : (\kappa \to t) \to e \to \kappa \to t$$

$$\llbracket \textit{know} \rrbracket = \lambda\phi, x, c. \begin{cases} \mathsf{know}_f(c)(\phi)(x) & \tau_{know}(c) \\ \mathsf{know}_{nf}(c)(\phi)(x) & \neg\tau_{know}(c) \end{cases}$$

We may additionally assume the following postulate, which ensures that the factive use of *know* entails the truth of the clause it selects at the context at which it is evaluated:

$$\forall c, \phi, x. \mathsf{know}_f(c)(\phi)(x) \to \phi(c)$$

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

Likewise for all clause-selecting predicates. Given a context c, those predicates which are always factive will have the meaning $\mathsf{verb}_f(c)$ with probability 1, and those which are never factive will have the meaning $\mathsf{verb}_{nf}(c)$ with probability 1.

It is important to note that while the lexical entry provided above for *know* may appear to render it semantically ambiguous, it in fact does not. Ambiguity, on our account, is a potential cause of metalinguistic uncertainty, but not of occasional uncertainty (ambiguities are resolved on particular occasions). Thus, whether the above entry for *know* renders it ambiguous versus, say, *vague* is a matter of how uncertainty about the value of the parameter τ_{know} is regulated; that is, whether its distribution is determined by metalinguistic uncertainty or occasional uncertainty.

We should point out that the kind of characterization of factivity we have presented here is *ad hoc*, in the sense that it provides no characterization of the general *factors* distinguishing different predicates' interpretations. This approach is ideal for current purposes, however, since our aim is not to provide an explanation for factivity, but to discover properties of its behavior, characterized at a level of abstraction that may be used to circumscribe more detailed acccounts.

In particular, we wish to determine whether the gradience it exhibits is a manifestation of occasional uncertainty (supporting the Fundamental Gradience Hypothesis) or metalinguistic uncertainty (supporting the Fundamental Discreteness Hypothesis). But that is all. In turn, a genuine account of factivity should effectively provide interpretations to the constants of our language, such as know_f and know_{cf}, which satisfy the postulate provided above. Put differently, we introduce just the right amount of abstraction to distinguish the Fundamental Discreteness Hypothesis from the Fundamental Gradience Hypothesis.

Note that alternatively, we could have described factivity as arising from a source external to the predicate; for example, by assuming that it is encoded by a complementizer (Kiparsky and Kiparsky 1970 *et seq.*).

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

Given such an alternative, we need not assume that clause-embedding predicates, such as *know*, themselves give rise to factive interpretations.[16]

$$\llbracket know \rrbracket : (\kappa \to t) \to e \to \kappa \to t$$
$$\llbracket know \rrbracket = \mathsf{know}$$

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;

[16] The implementation according to 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 Kratzer 2006; Moulton 2009; Bogal-Allbritten 2016; White and Rawlins 2018a) and may be adapted to other such approaches that regard the embedded clause as an intersective modifier of eventualities (Elliott 2016; Elliott 2020).

Altmann and Kamide 1999, i.a.). If the context makes such information available, this implementation can then yield the same range of statistical models as one which assumes that factivity is driven principally by the lexical entries for individual predicates.

4 Modeling

To investigate the theories of factivity and world knowledge which are 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 model 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. In Section 4.2, we describe and compare the two models which we fit to Degen and Tonhauser's norming experiment data: a gradient model and a discrete model. We then describe each of the models of factivity possible under our framework in Section 4.3 before reporting our comparisons of these models, given Degen and Tonhauser's projection experiment data, in Section 4.4.

4.1 Linking to response behavior

To connect the probabilistic programs associated with sentences to actual data, we need linking assumptions that relate these programs to the inference judgments that experimental participants report on a slider scale going between *no* and *yes* (cf. Jasbi, Waldon, and Degen 2019).

We assume that a slider scale can take on values in the closed unit interval $[0, 1]$. Thus, specifying linking assumptions requires saying how probabilistic programs may be used to compute distributions on this interval.

4.1.1 Response models in the abstract

We specify our models abstractly by defining a class of functions $\mathtt{respond}^{f,\Psi}_{(\cdot)}$, each of which takes a probabilistic program m of type $\mathtt{P}\kappa$ that computes a distribution over contexts, as well as a possible inference ϕ of type $\kappa \longrightarrow t$, and then associates them with a distribution over slider responses on the closed unit interval:

$$\begin{aligned} \mathtt{respond}^{f,\Psi}_{(\cdot)} &\;:\; \mathtt{P}\kappa \longrightarrow (\kappa \longrightarrow t) \longrightarrow \mathtt{P}r \\ \mathtt{respond}^{f,\Psi}_{c\sim m}(\phi(c)) &\;=\; \begin{array}{l} x \sim \boxed{\mathbb{P}\left(\begin{array}{c} c \sim m \\ \boxed{\phi(c)} \end{array}\right)} \\ f(x, \Psi) \end{array} \end{aligned}$$

This model 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 in $[0, 1]$ according to the probabilistic program $f(x, \Psi)$ for some fixed family of distributions $f : r \times p_1 \times ... \times p_n \rightarrow \mathsf{P}r$ and vector $\Psi : p_1 \times ... \times p_n$ of nuisance parameters.

To give a schematic example, let's say that the common ground of interest is characterized by a probabilistic program $\mathtt{commonGround}$ of type $\mathsf{P}(\mathsf{P}\kappa)$. Then the following program of type $\mathsf{P}r$ characterizes the distribution of slider responses on the closed 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*:

$$\begin{aligned}
& m \sim \mathtt{commonGround} \\
& m' \sim \boxed{\mathtt{update}(\llbracket \textit{Susan knows that Grace visited her sister} \rrbracket)(m)} \\
& \mathtt{respond}^{f,\Psi}_{c \sim m'}(\llbracket \textit{Grace visited her sister} \rrbracket^c) \\
= \quad & m \sim \mathtt{commonGround} \\
& x \sim \boxed{\mathbb{P}\left(\begin{array}{l} c \sim m \\ \mathtt{observe}(\llbracket \textit{Susan knows that Grace visited her sister} \rrbracket^c) \\ \boxed{\llbracket \textit{Grace visited her sister} \rrbracket^c} \end{array} \right)} \\
& f(x, \Psi)
\end{aligned}$$

This probabilistic program first samples a probabilistic program m of type $\mathsf{P}\kappa$ that computes a distribution over contexts. Under the distribution represented by m, the parameters regulating metalinguistic uncertainty have been *fixed*, but the parameters regulating occasional 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 c from m, which it uses to perform Bayesian update—by observing that *Susan knows that Grace visited her sister* is true, given c; before (ii) returning T or F, depending on which is the value of the interpretation of *Grace visted her sister*, given c.

4.1.2 Examples of response models

For the purpose of specifying our models more concretely, 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 $\boldsymbol{\tau_w}$ determining whether or not each fact under consideration related to world knowledge is true or false, and the inhabitants of t^n are n-tuples of truth values $\boldsymbol{\tau_v}$ determining whether or not the complement of each predicate under consideration indeed *projects*. Thus, each model is of the form:

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

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:

$$\begin{array}{l} m \sim \texttt{commonGround} \\ x \sim \boxed{\mathbb{P}\left(\begin{array}{l} \langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle \sim m \\ \texttt{observe}(\llbracket \textit{Susan knows that Grace visited her sister} \rrbracket^{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle}) \\ \boxed{\llbracket \textit{Grace visited her sister} \rrbracket^{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle}} \end{array} \right)} \\ f(x, \Psi) \end{array}$$

Moreover, if we take $\mathsf{know}(\boldsymbol{\tau_v})$ to be the component of $\boldsymbol{\tau_v}$ that says whether or not the complement of *know* projects, and $\mathsf{grace}(\boldsymbol{\tau_w})$ to be the component of $\boldsymbol{\tau_w}$ that says whether or not Grace visited her sister, then the program above can be rephrased as follows:[17]

$$\begin{array}{l} m \sim \texttt{commonGround} \\ x \sim \boxed{\mathbb{P}\left(\begin{array}{l} \langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle \sim m \\ \boxed{\mathsf{know}(\boldsymbol{\tau_v}) \vee \mathsf{grace}(\boldsymbol{\tau_w})} \end{array} \right)} \\ f(x, \Psi) \end{array}$$

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 $\mathsf{know}(\boldsymbol{\tau_v})$ to be T) or for it to be entailed by prior knowledge (i.e., for $\mathsf{grace}(\boldsymbol{\tau_w})$ to be T), its semantic value is equivalent to a disjunction.[18]

The general set-up described here will remain invariant under the models 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 occasional uncertainty.

The remaining challenge in implementing our proposal is to find a family of distributions over $[0, 1]$ that can flexibly capture a variety of possible response distributions. The reason finding such a family is challenging is that many of the families of bounded distributions commonly used in generalized linear mixed models, such as the beta or logit-normal distributions, only have support on the open interval $(0, 1)$; that is, beta and logit-normal random variables never take on 0 or 1 values. As a result, they cannot describe the distribution of slider responses. This fact is particularly problematic in the current context, where endpoint responses are meaningful by hypothesis.[19]

We consider two approaches to this challenge. The first is *truncation*, whereby we model distributions on the closed unit interval by *truncating* a distribution on $\mathbb{R}$. The second is *inflation*, whereby we model 0 and 1 responses as arising from a response process that mixes a distribution on $(0, 1)$ with a distribution on $\{0, 1\}$, thereby *inflating* the number of 0's and 1's relative to that expected under the distribution on $(0, 1)$. To implement the truncation approach, we assume f

[17] Strictly speaking, we prove this equivalence only for the models which take world knowledge to give rise to occasional uncertainty (see Appendix C.1.2). Note that this assumption is *a priori* justified by the results we obtain from modeling the norming data (see Section 4.2.3), though for completeness, we also fit models of factivity which consider world knowledge to be discrete.

[18] The full model specifications provided in Appendix C make use of this fact.

[19] A common approach used in regression analyses is to sidestep this issue by "nudging" 0 and 1 responses toward 0.5 by a small δ, thereby putting them into $(0, 1)$. This approach is itself problematic because it introduces an additional unnecessary research degree-of-freedom; moreover, it is especially problematic in the current context for reasons which we discuss in Appendix A.

to be the family of truncated normal distributions $\mathcal{N}(\cdot)\,\mathtt{T}[0,1]$, with nuisance parameter $\Psi \equiv \sigma$. To implement the inflation approach, we assume f to be the family of ordered beta distributions $\mathtt{OrdBeta}(\cdot)$, with nuisance parameters $\Psi \equiv \langle \phi, \mathbf{c} \rangle$ (Kubinec 2023).[20] Ultimately, we find that our main results are insensitive to which approach we take, so we focus on models implementing the truncation approach here (see Appendix B and Appendix C.2 for discussion of the inflation models).

4.2 Models of the norming data

To construct our models, we use a *pipelined* approach:[21] we first fit models 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*.[22] We then used these posterior distributions as prior distributions for the corresponding parameters in the four models of Degen and Tonhauser's projection experiment data (experiment 2b) which are implied by the framework described in Section 3. We present these models in Section 4.3.

Recall that the prompts with which participants were presented in the norming task inquired about the likelihood of a fact related to world knowledge, given some other background fact. To model such inference judgments, we fit two kinds of models: (i) a *gradient* model, according to which inferences conditioned by world knowledge are gradient in nature; and (ii) a *discrete* model, according to which such inferences are discrete in nature. It turns out that the gradient model of the norming data provides the better fit (see Section 4.2.3). We therefore use the gradient model to extract prior distributions for parameters of the factivity models.

4.2.1 World knowledge as a gradient phenomenon

The gradient model of the norming data encodes uncertainty about world knowledge as occasional uncertainty. For expository purposes, we ignore the distributions over parameters related to factivity (i.e., we simply replace the irrelevant parameters $\boldsymbol{v}$ and $\boldsymbol{\tau_v}$ by underscores and leave out the latter's sampling statement). The corresponding model of the common ground may then be given as follows:

$$\begin{array}{rcl} \texttt{norming-gradient} & : & \mathsf{P}(\mathsf{P}\kappa) \\ \texttt{norming-gradient} & = & \langle \boldsymbol{w}, \_ \rangle \sim \texttt{priors} \\ & & \boxed{\begin{array}{l} \boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\ \boxed{\langle \boldsymbol{\tau_w}, \_ \rangle} \end{array}} \end{array}$$

Importantly, the sampling statement $\boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w})$ which determines the values of parameters regulating world knowledge occurs *inside* the outer orange box; thus the expected values of these parameters will end up as the values targeted by the response.

[20] See Appendix B for details on the ordered beta distribution.

[21] We provide additional formal details about the modeling pipeline in Appendix C.

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

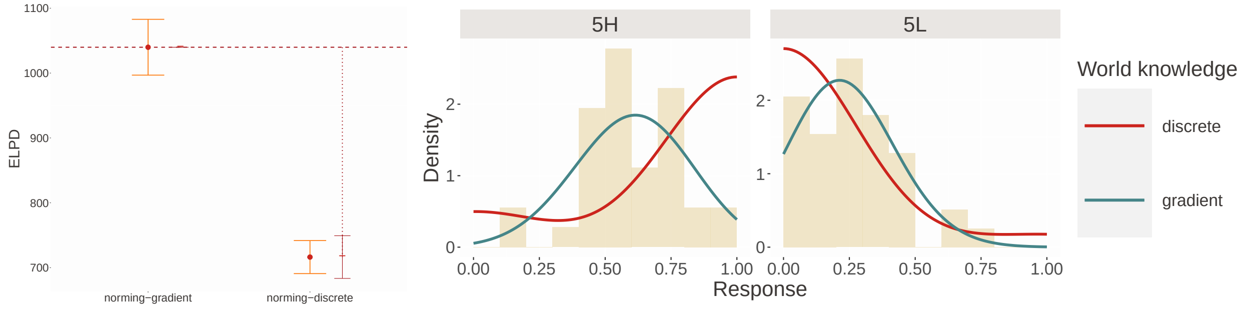


Figure 7: Left: ELPDs for the two models of the norming data. Dotted line indicates estimated difference between the norming-discrete model and the norming-gradient model. Error bars indicate standard errors. Right: posterior predictive distributions for both models for the item *Sophia got a tattoo*, given either the fact *Sophia is a hipster* (5H) or the fact *Sophia is a fashion model* (5L).

4.2.2 World knowledge as a discrete phenomenon

The discrete model of the norming data encodes uncertainty about world knowledge as type-level uncertainty. It is given as follows:

$$\texttt{norming-discrete} : \mathtt{P}(\mathtt{P}\kappa)$$

$$\begin{aligned}
\texttt{norming-discrete} \;=\; & \langle \boldsymbol{w}, \_ \rangle \sim \texttt{priors} \\
& \boldsymbol{\tau}_{\boldsymbol{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\
& \boxed{\langle \boldsymbol{\tau}_{\boldsymbol{w}}, \_ \rangle}
\end{aligned}$$

The only difference between this model and the gradient model is the location of the sampling statement for the parameters encoding world knowledge. According to the discrete model, world knowledge parameters are sampled once on each response occasion, rather than averaged over.

4.2.3 Comparison

The ELPDs for the two models of the norming data are shown in Figure 7 (left plot). It can be seen that the gradient model substantially outperforms the discrete model, indicating that responders represent their uncertainty about world knowledge gradiently.

To give a sense of why the gradient model performs so much better, Figure 7 also shows the posterior predictive distributions for the two models for two items (right plot; see Figure 18 of Appendix D.2 for all items). These plots provide a visual indicator of how well each model fits the distribution of responses to the norming items: the closer a particular curve is to the shape of the histogram, the better the corresponding model fits the data. Thus, the gradient model fits the data better due to the fact that responses often cluster away from the endpoints of the scale. Meanwhile, whereas the gradient model can capture this behavior, the discrete model only allows the theoretical response (i.e., before any error is added) to occur at a scale endpoint. This limitation often causes the posterior predictive of the discrete model to have a bimodal distribution, as the plot for 5H illustrates.

4.3 Models of factivity

Now we provide our four models of factivity and prior world knowledge, which we fit to Degen and Tonhauser's projection experiment data.[23] 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. The possibility that predicates vary in whether the factivity inferences they license are discrete or gradient is an interesting one, and we investigate it in ongoing work.

4.3.1 Factivity as a fundamentally discrete phenomenon

The first theory we consider regards sentences including clause-embedding predicates as either triggering or not triggering factive inferences on particular occasions of use, depending on either (i) the interpretation of (a) the clause-embedding predicate, or (b) the surrounding structure; or (ii) some discourse structure, such a the common ground or question under discussion, whose properties are conditioned on pragmatic knowledge about the sentence. The uncertainty about whether or not a given predicate's complement clause projects is thus regarded as metalinguistic in nature.

Assuming factivity is a discrete phenomenon, the question remains how world knowledge might affect veridicality inferences about the content of an embedded clause. Here two possibilities arise, depending on the scopes of the parameters $\boldsymbol{\tau_w}$ and $\boldsymbol{\tau_v}$. We discuss these in turn.

The first possibility allows for uncertainty related to world knowledge to manifest itself as occasional 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:

$$\begin{aligned}
\texttt{discrete-factivity} &: \mathsf{P}(\mathsf{P}\kappa)\\
\texttt{discrete-factivity} &= \langle \boldsymbol{w}, \boldsymbol{v}\rangle \sim \texttt{priors}\\
&\ \boldsymbol{\tau_v} \sim \texttt{Bernoulli}(\boldsymbol{v})\\
&\ \boxed{\begin{array}{l}\boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w})\\ \boxed{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle}\end{array}}
\end{aligned}$$

The aspect of this model crucial to the way in which it regards factivity is the location of the sampling statement $\boldsymbol{\tau_v} \sim \texttt{Bernoulli}(\boldsymbol{v})$; in particular, it is crucial that this statement occurs *prior* to returning the probabilistic program of type $\mathsf{P}\kappa$ that characterizes occasional uncertainty—that is, outside of the orange boxes. As a result, whether or not the complement of a given predicate projects is fixed on individual occasions of interpretation. By contrast, the sampling statement $\boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w})$ is part of the returned program, rendering uncertainty about world knowledge occasional (token-level) in nature.

[23] See Appendix C for the full model specifications.

The second possibility is that both factivity and world knowledge are discrete in nature.

$$\begin{array}{rl}\texttt{wholly-discrete} : \mathsf{P}(\mathsf{P}\kappa) & \\ \texttt{wholly-discrete} = & \langle \boldsymbol{w}, \boldsymbol{v} \rangle \sim \texttt{priors} \\ & \boldsymbol{\tau_v} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\ & \boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\ & \boxed{\boxed{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle}}\end{array}$$

This *wholly-discrete* model effectively hypothesizes that no inferences display uncertainty on particular occasions of use.[24] Under this version, the sampling statement $\boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w})$ has also been moved outside of the outer orange box. Any gradience displayed in people's measured inferences must therefore be due to task effects.

Note that the `discrete-factivity` model effectively provides a more complete version of the `norming-gradient` model described in Section 4.2 by making a decision about where to sample the parameters $\boldsymbol{\tau_v}$. Likewise, the `wholly-discrete` model effectively completes the `norming-discrete` model. We therefore might have an *a priori* expectation that of the two, the `discrete-factivity` model should provide a better fit to the data.

4.3.2 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 occasional uncertainty level. This theory also has two possible implementations that vary with respect to whether world knowledge is modeled as token-level uncertain or metalinguistically uncertain. We refer to the version of the theory that regards both factivity and world knowledge as token-level uncertain as the *wholly-gradient* version.

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:

$$\begin{array}{rl}\texttt{wholly-gradient} : \mathsf{P}(\mathsf{P}\kappa) & \\ \texttt{wholly-gradient} = & \langle \boldsymbol{w}, \boldsymbol{v} \rangle \sim \texttt{priors} \\ & \boxed{\begin{array}{l}\boldsymbol{\tau_v} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\ \boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\ \boxed{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v} \rangle}\end{array}}\end{array}$$

We refer to the alternative version of the model that regards world knowledge as discrete and factivity as gradient as the *discrete-world* version. As with the wholly-discrete model, we do not take anyone's theory of factivity to be consistent with the discrete-world model, in particular, but we include it, since it is a possibility implied by our framework.[25] According to this version, the locations of the sampling statements which were used to encode the discrete-factivity model are

[24] We do not take anyone to endorse such a model; and indeed, the results from our fits to the norming data speak against it. We include it mainly because it is a logical possibility under our framework.

[25] Again, the results from our fits to the norming data speak against the discrete-world model, and we don't take anyone to endorse it. We include it merely because it is a logical possibility under our framework.

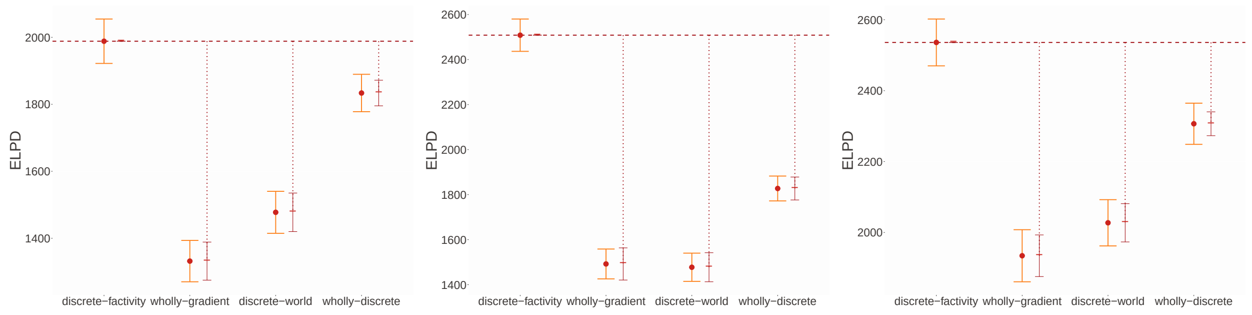


Figure 8: Left: ELPDs for the four models. Center: ELPDs for the four models with an anti-veridicality component added for each predicate. Right: ELPDs for the four model evaluations on our replication experiment data. Dotted lines indicate estimated differences between each model and the discrete-factivity model. Error bars indicate standard errors.

switched:

$$\begin{aligned}
\texttt{discrete-world} &: \mathsf{P}(\mathsf{P}\kappa)\\
\texttt{discrete-world} &= \langle \boldsymbol{w}, \boldsymbol{v}\rangle \sim \texttt{priors}\\
&\ \boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w})\\
&\ \boxed{\begin{array}{l} \boldsymbol{\tau_v} \sim \texttt{Bernoulli}(\boldsymbol{v})\\ \boxed{\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle}\end{array}}
\end{aligned}$$

Among the four models we consider, we take the `wholly-gradient` model to be the one that comes closest to implementing Tonhauser, Beaver, and Degen's (2018) proposal. Note that this model effectively provides an alternative completion of the `norming-gradient` model described in Section 4.2; meanwhile, the `discrete-world` model provides an alternative completion of the `norming-discrete` model.

4.4 Comparisons

Figure 8 (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; 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. Further, note that 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 9 (see Figure 19 of Appendix D.2 for all predicates).

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

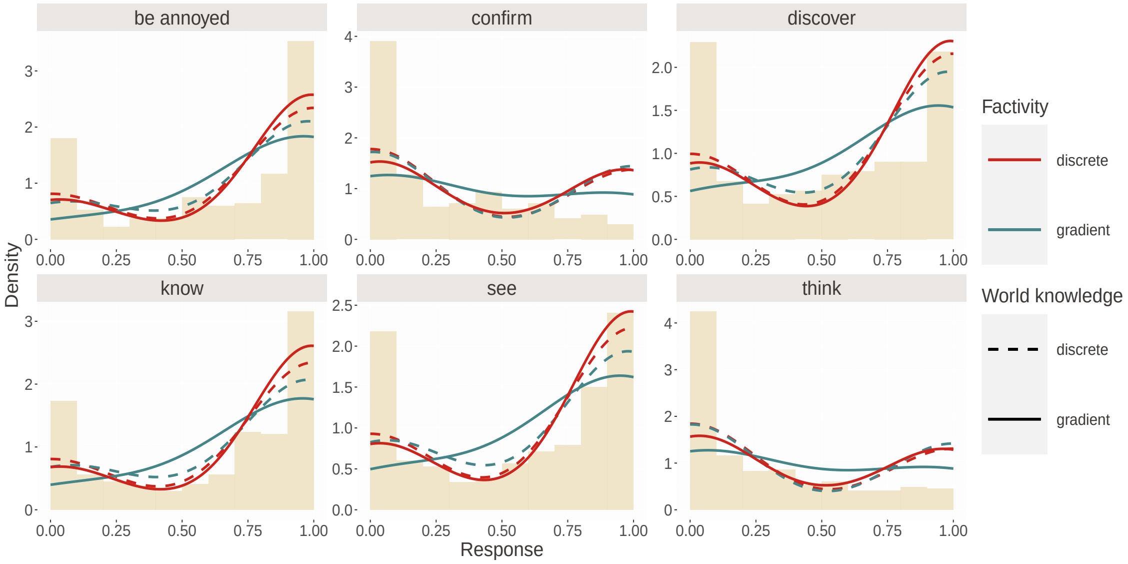


Figure 9: Posterior predictive distributions (with simulated participant intercepts) of all four models for six predicates from Degen and Tonhauser's (2021) projection experiment, for all contexts combined. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

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 of the models.

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.[26] This property of the models that assume some amount of discreteness makes them more complex than the wholly-gradient model—in the sense that they have more effective parameters—but as the pattern of ELPDs in Figure 8 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 19. This difficulty appears to be due to an anti-veridicality inference associated with these predicates; i.e., an inference that the content of the complement clause is *not* true.[27]

[26] Multi-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 9. Such multi-modality can only arise due to participant random effects. See Appendix C for formal details concerning these random effects.

[27] These 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

Since we consider denotations for predicates that vary only with respect to veridicality and non-veridicality—not anti-veridicality—no model captures the latter. To assume a three-way distinction among veridicality, non-veridicality, and anti-veridicality would effectively be to allow the discrete-factivity and wholly-discrete models to mix in another distribution with a mode at 0; meanwhile, it would effectively allow the wholly-gradient model an extra parameter to modulate the gradience associated with individual predicates.

To investigate the possibility that such a modification might affect the model ranking, we additionally fit models which add an anti-veridicality component for each predicate (see Appendix C.1.2). We find that, while the anti-veridicality component substantially improves the performance of the discrete-factivity model (as illustrated by the plots in Figure 20 of Appendix D) the only change in relative model performance is that the wholly-gradient model now slightly outperforms the discrete-world model, as shown in Figure 8 (center plot).

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.

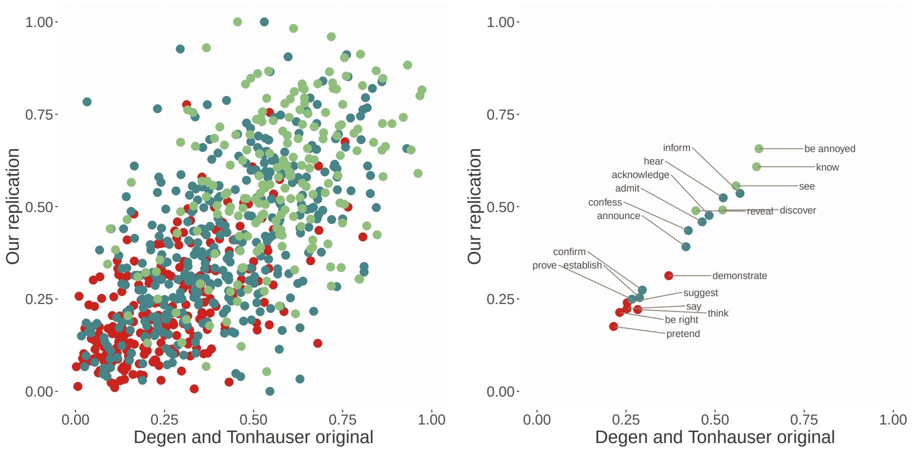


Figure 10: Degen and Tonhauser's (2021) projection data versus our replication. Left: item means (Spearman's $r = 0.68, p \leq 0.001$). Right: verb means (Spearman's $r = 0.98, p \leq 0.001$). For both, "non-factive" verbs are in red, "optionally factive" verbs are in teal, and "canonically factive" verbs are in green.

5.1.3 Results

Figure 10 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 $r = 0.98, p \leq 0.001$) 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 $\boldsymbol{\mu}_{\nu}$ and standard deviations $\boldsymbol{\sigma}_{\nu}$ of the marginal posterior distribution of the log-odds of projection for each predicate, as well as means $\boldsymbol{\mu}_{\omega}$ and standard deviations $\boldsymbol{\sigma}_{\omega}$ of the marginal posterior distribution of the log-odds certainty for each context. We then use normal distributions with these means and standard deviations as prior distributions for the corresponding parameters in the models constructed for the evaluations.[28]

5.1.5 Model comparison

Figure 8 (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 8: the discrete-factivity model fares the best, followed by the wholly-discrete model. The wholly-gradient and discrete-world models fare the worst.

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.

[28] See Appendix C.1.3 for further details concerning these models.

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.

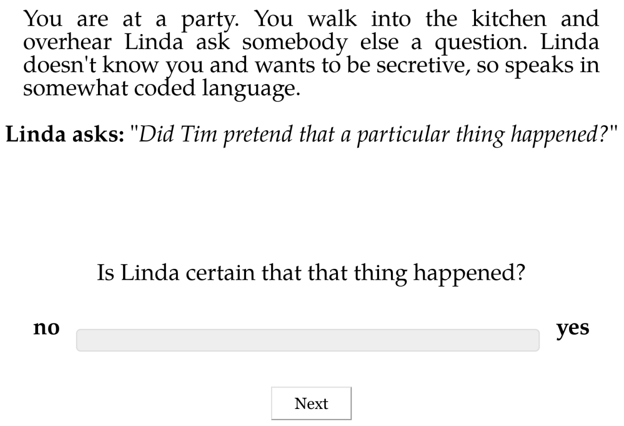
5.2 Experiments 2 and 3: non-contentful contexts

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.*

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.

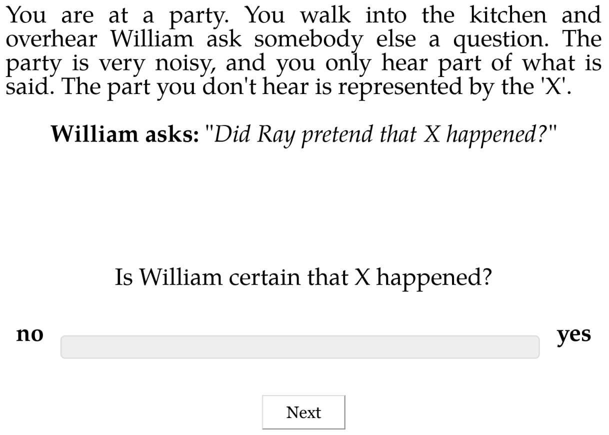
Linda asks: "*Did Tim pretend that a particular thing happened?*"

Is Linda certain that that thing happened?

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 experiment, 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?

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?*.

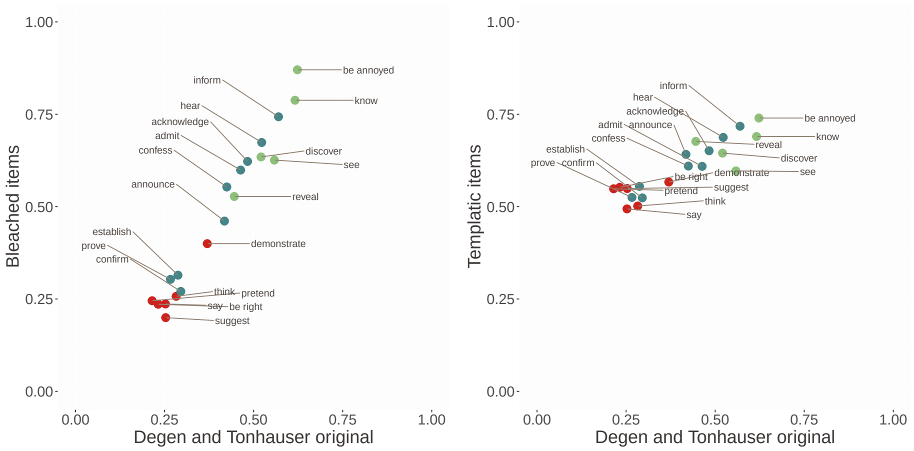


Figure 11: Degen and Tonhauser's (2021) projection data versus data from contexts with minimal lexical content. Left: bleached data (Spearman's $r = 0.97, p \leq 0.001$). Right: templatic data (Spearman's $r = 0.87, p \leq 0.001$). For both, "non-factive" verbs are in red, "optionally factive" verbs are in teal, and "canonically factive" verbs are in green.

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 11 that the responses elicited by both the bleached (Spearman's $r = 0.97, p \leq 0.001$) and templatic (Spearman's $r = 0.87, p \leq 0.001$) 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.[29] This 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.

[29] Notably, 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.

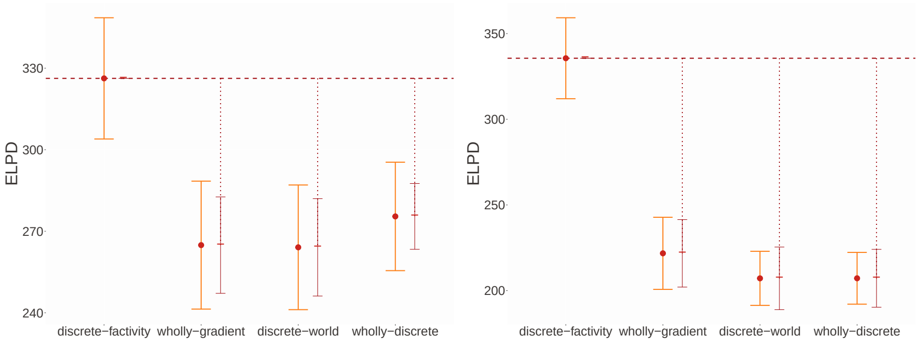


Figure 12: ELPDs for the four model evaluations on the bleached data (left) and the templatic data (right). Dotted lines indicate estimated differences between each model and the discrete-factivity model. Error bars indicate standard errors.

5.2.5 Model fitting

To evaluate the four models using both the bleached and templatic data, we use the same means $\boldsymbol{\mu}_\nu$ and standard deviations $\boldsymbol{\sigma}_\nu$ 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.[30]

5.2.6 Model comparison

Figure 12 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 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 11. 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 21 and 22 of Appendix D.2 show the posterior predictive distributions of these evaluations for all predicates.

[30] See Appendix C.1.4 for further details concerning these models.

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. First, 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 second, whether they consider the contribution of the relevant predicate itself to these inferences to be gradient or discrete. We refer to a given factor as "gradient" if varying that factor produces a continuous effect on the magnitude of the judgment associated with the inference; and we refer to it as "discrete" if varying the factor affects the probability with which the inference is judged as certain, as opposed to remaining 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; see Figure 8, left plot). 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 8, 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 12).

Taken together, these results provide strong evidence that the gradience among the clause-embedding predicates observed by White and Rawlins (2018b) and studied by Degen and Tonhauser is *metalinguistic*; i.e., it has a *type-level* source. Clause-embedding predicates differ in the frequencies with which they trigger projective inferences, but the contribution a predicate makes on particular occasions of use (or interpretation) is *discrete*: either it produces the relevant inference, or it does not produce it at all. Thus, these results provide strong support for the Fundamental Discreteness Hypothesis, discussed in Section 2. What do these findings entail for theories of factivity, in particular, and presupposition projection, as a whole?

In Section 2, we considered three ways in which the Fundamental Discreteness Hypothesis might be cashed out: (i) clause-embedding predicates may have multiple senses—at least one that is implicated in triggering projection and at least one that is not; (ii) clause-embedding predicates may occur in multiple structures—at least one that is implicated in triggering projection and at least one that is not; or (iii) clause-embedding predicates may be usable in contexts where the common ground—or some other construct, such as the question under discussion—entails the content of its embedded clause, as well as contexts where it does not (see Simons, Beaver, et al. 2017; Roberts and Simons to appear). Under the first two explanations, which we describe as *conventionalist*, the metalinguistic uncertainty observed is uncertainty about how lexical or structural ambiguity tends to be resolved; under the third explanation, which we describe as *conversationalist*, the metalinguistic uncertainty observed is uncertainty about what kind of discourse contexts an expression tends to occur in.

How might we empirically distinguish these types of explanation? A crucial difference between conventionalist accounts, on the one hand, and conversationalist accounts, on the other, is that the former deem semantic knowledge about an expression to be necessary (though perhaps not sufficient) to trigger a projective inference, while the latter do not.

6.1 Conventionalist accounts

Consider an emotive predicate such as *be annoyed*. In addition to giving rise to veridicality inferences, such predicates tend to give rise to inferences about the beliefs of the entity they are predicated of. For example, (12) tends to give rise to both the inferences in (13a) and in (13b).

(12) Was Jo annoyed that Mo left?

(13) a. Jo believed that Mo left.

b. Mo left.

One species of conventionalist account attributes (13a) and (13b) to independent semantic properties of (12) (see Villalta 2008; Romero 2015; Uegaki and Sudo 2019, i.a.).

Another species attributes both inferences to the same property—e.g., deriving one from the other by invoking additional (presumably pragmatic) assumptions (Djärv, Zehr, and Schwarz 2018). For instance, adapting ideas put forth by Heim (1992, p. 212), (13b) might be derived from (13a) by invoking a defeasible assumption that Jo is credible. On this sort of account, (13a) and (13b) need not be triggered independently.

This second species of account can be extended to emotive factives as a class (Djärv, Zehr, and Schwarz 2018, p. 382), given that all emotive predicates tend to give rise to such belief inferences (see Kane, Gantt, and White 2022). Gradience among the observed inference judgments might then be a product of variability either in whether or not a belief inference is drawn or in whether or not an assumption of credibility is made.

6.2 Conversationalist accounts

In contrast, conversationalist approaches posit that knowledge of predicates like *be annoyed* relates those predicates to distributions over discourse-theoretic constructs (or classes thereof), such as the common ground or the question under discussion. Such distributions are presumably constrained by general pragmatic principles, which may, in turn, interact with conceptual knowledge associated with a given predicate (see Simons, Beaver, et al. 2017; Roberts and Simons to appear). Crucially, lexical semantic knowledge is not the main driver of projective inferences under this sort of account—if it is implicated at all.

For instance, Simons, Beaver, et al. put forth a conversationalist account that relies on the notion of a QUD. On their account, whether or not the complement of a clause-embedding predicate projects varies according to prosodic and contextual factors conditioned by the QUD.[31] Specifically, Simons, Beaver, et al. argue that the projective inferences licensed at any point in a discourse are simply those which are backgrounded by the QUD, while the non-projective inferences are those which are made at-issue (see also Simons 2001; Simons 2007, i.a.).

[31] Importantly, lexical knowledge also conditions QUD choice under standard dynamic accounts in the Karttunen-Heim tradition; it just does so indirectly. If, following Heim (1983), a sentence denotes a context-change potential, and its presuppositions are encoded as definedness conditions, then it will only be defined on context sets that satisfy its presuppositions. Since a QUD is just a partition of the context set (Roberts 2012), the update denoted by a sentence is defined just in case that sentence's presuppositions are true in every cell of the partition—i.e., if the QUD entails the sentence's presuppositions. Metalinguistic uncertainty about a sentence's definedness conditions can therefore result in the defeasibility of inferences that are driven by those definedness conditions.

Roberts and Simons (to appear) extend this idea by arguing that "projectivity... is predictable from the detailed lexical semantics of the triggers, in combination with broad pragmatic principles" (p. 2), but crucially also that "projective readings of [change of state] predicates, of factives, and of predicates that trigger selectional restrictions, are not conventionally derived" (p. 29). Rather, they hold that these projective inferences are driven by "the nuanced and probabilistic knowledge that language users have about the functions, effects and contexts of use of linguistic forms" (p. 27), and they work through various specific examples.

6.3 Distinguishing conventionalist from conversationalist accounts

Since conventionalist and conversationalist accounts diverge in how they regard the relationship between projection and semantic knowledge, they potentially make different predictions about whether or not a predicate's (degree of) factivity should correlate with its syntactic distribution. Conventionalist accounts make the (relatively strong) prediction that, because projective inferences are conditioned by semantic properties of predicates, their presence should correlate with whatever aspects of a predicate's syntactic distribution those semantic properties correlate with themselves. Conversationalist accounts make no such commitments. Indeed, insofar as projection is "a function of a complex set of factors, some lexical, others contextual" (Roberts and Simons to appear, p. 53), these accounts may predict very little if any correlation between factivity and distribution. How strong of a correlation any particular account does, in fact, predict is highly dependent on how strong that account regards the influence of distributionally correlated lexical semantic properties to be on a predicate's distribution of co-occurrences with discourse-theoretic constructs (or classes thereof), relative to other factors.

In attempting to distinguish conventionalist accounts from conversationalist accounts, it may therefore be useful to ask whether or not we find strong correlations between predicates' projective inferences and their distributional properties. As it happens, we do. Recall that in Section 2, we discussed whether or not there are in fact classes of factive predicates, and we noted that Kane, Gantt, and White (2022) establish the existence of multiple factive subclasses. Given its relevance to the question at hand, the way Kane, Gantt, and White derive their clustering is important: they select the clustering that optimally predicts the acceptability of predicates in different subcategorization frames. They find that the optimal clustering is able to predict acceptability at a level comparable to—in fact, slightly better than—that reported by White and Rawlins (2020) for models predicting a predicate's acceptability in different subcategorization frames from its corpus frequency in those frames.

This strong correlation may be evidence either (i) for a conventionalist account; or (ii) for a conversationalist account that relies relatively more heavily on lexical semantic factors over other factors in its account of how distributions over discourse-theoretic constructs are conditioned. What seem less likely to be consistent with these findings are conversationalist accounts that put a heavy load on contextual factors as drivers of projective inferences, compared to distributionally correlated lexical semantic factors.

7 Conclusion

The results reported here can be understood as broadly consistent with a fairly traditional view of factivity, in line with what was originally advocated by Kiparsky and Kiparsky (1970), Karttunen (1971), inter alia. According to this view, some predicates can be understood to trigger a presupposition that the clause they select is true. The key departure from this tradition which we would advocate, based on our results (and following Degen and Tonhauser), is in the particular classification of predicates that researchers ought to appeal to. Indeed, *none* of the predicates that Degen and Tonhauser investigate appears to be assigned a factive interpretation in all of its uses; rather, all predicates are 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, uncertainty about the semantic properties of linguistic expressions will be an essential feature of that knowledge.

References

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.

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 Crni\v{c} 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.

Barwise, Jon and John Perry. 1983. *Situations and attitudes*. Cambridge: MIT Press.

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.

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.

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.

Bogal-Allbritten, Elizabeth A. 2016. Building Meaning in Navajo. PhD thesis. University of Massachusetts, Amherst.

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.

Coppock, Elizabeth. 2018. Outlook-based semantics. *Linguistics and Philosophy* 41.2, pp. 125–164. DOI: 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.

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.

Djärv, 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.

Djärv, 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.

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.

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.

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. $\langle M, \eta, \star \rangle$ 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.

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: 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.

Grice, Paul. 1989. *Studies in the way of words.* Cambridge: Harvard University Press.

Grove, Julian. 2022. Presupposition projection as a scope phenomenon. *Semantics and Pragmatics* 15, 15:1–39. DOI: 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.

Heim, Irene. 1982. The Semantics of Definite and Indefinite Noun Phrases. PhD Thesis. Amherst: University of Massachussetts.

Heim, Irene. 1983. On the Projection Problem for Presuppositions. *The West Coast Conference on Formal Linguistics (WCCFL).* Ed. by Michael D. Barlow, Daniel P. Flickinger, and Michael Westcoat. Vol. 2. Stanford: Stanford University Press, pp. 114–125.

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.

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.

Jasbi, Masoud, Brandon Waldon, and Judith Degen. 2019. Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate. *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.00189.

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.

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.

Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4.1, pp. 55–69. DOI: 10.1080/08351817109370248.

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.

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.

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.

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.

Kiparsky, Paul and Carol Kiparsky. 1970. FACT. In *Progress in Linguistics*. De Gruyter Mouton, pp. 143–173.

Kratzer, Angelika. 2006. *Decomposing attitude verbs*. The Hebrew University of Jerusalem.

Kubinec, Robert. 2023. Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds. *Political Analysis* 31.4, pp. 519–536. DOI: 10.1017/pan.2022.20.

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.

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.

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.

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.

Moulton, Keir. 2009. Natural Selection and the Syntax of Clausal Complementation. PhD thesis. University of Massachusetts, Amherst.

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.

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.

Roberts, Craige and Mandy Simons. to appear. Preconditions and Projection: Explaining Non-Anaphoric Presupposition. *Linguistics and Philosophy*.

Romero, Maribel. 2015. Surprise-Predicates, Strong Exhaustivity and Alternative Questions. *Semantics and Linguistic Theory*, pp. 225–245. DOI: 10.3765/salt.v25i0.3081.

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.

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. 2001. On the conversational basis of some presuppositions. *Semantics and Linguistic Theory* 11. Ed. by R. Hasting, B. Jackson, and Z. Zvolensky, pp. 431–448.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117.6, pp. 1034–1056. DOI: 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.

Stan Development Team. 2023. *Stan Modeling Language Users Guide and Reference Manual, 2.32*. Tech. rep.

Tonhauser, Judith. 2016. Prosodic cues to presupposition projection. *Semantics and Linguistic Theory* 26.0, pp. 934–960. DOI: 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.

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.

Uegaki, Wataru and Yasutada Sudo. 2019. The *hope-wh puzzle. *Natural Language Semantics* 27.4, pp. 323–356. DOI: 10.1007/s11050-019-09156-5.

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.

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*.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31.4, pp. 467–522. DOI: 10.1007/s10988-008-9046-x.

Von Fintel, Kai and Irene Heim. 2021. Intensional semantics. MIT.

Waismann, Friedrich. 1945. Verifiability. *Proceedings of the Aristotelian Society, Supplementary Volume*. Vol. 19, pp. 119–150.

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.

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.

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.

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 and Kyle Rawlins. 2020. Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: a journal of general linguistics* 5.1. DOI: 10.5334/gjgl.1001.

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.

A Beta distributions as likelihoods

Beta distributions cannot be used to model slider scales allowing endpoint responses. One common approach used in regression analyses sidesteps this issue by “nudging” 0 and 1 responses toward 0.5 by a small δ, thereby including them in $(0, 1)$. This approach is problematic because it introduces an unnecessary researcher degree of freedom—that of deciding by what amount δ to nudge the 0’s and 1’s inward—which can be used to manipulate model comparison metrics, and which we have no principled way of setting *a priori*.

First, note that WAIC relies crucially on the likelihood of the model. Second, note that the beta likelihood is highly sensitive to how much one nudges 0’s and 1’s inward, and that this sensitivity increases with the number of 0's and 1's present in the data. To see this, recall that the beta likelihood is:

$$\mathrm{Beta}(\mathbf{x} \mid \alpha, \beta) = \frac{\prod_i x_i^{\alpha-1}(1-x_i)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

The log-likeihood is thus:

$$\mathcal{L}_{\mathrm{Beta}}(\alpha, \beta \mid \mathbf{x}) = -\log \mathrm{B}(\alpha, \beta) + \sum_i (\alpha - 1)\log x_i + (\beta - 1)\log(1 - x_i)$$

Now suppose that some of the $x_i \in \{0, 1\}$, and therefore that the beta distribution would be undefined, since we would be taking $\log 0$. Let's say we nudge these x_i inward by $\delta < 0.01$ to avoid this outcome. Then we would obtain the likelihood:

$$\begin{aligned}\mathcal{L}(\alpha, \beta \mid \mathbf{x}) = -\log \mathrm{B}(\alpha, \beta) &+ \sum_{i:x_i=1} (\alpha - 1)\log(1 - \delta) + (\beta - 1)\log \delta \\ &+ \sum_{i:x_i=0} (\alpha - 1)\log \delta + (\beta - 1)\log(1 - \delta) \\ &+ \sum_{i:x_i \notin \{0,1\}} (\alpha - 1)\log x_i + (\beta - 1)\log(1 - x_i)\end{aligned}$$

Thus when $x_i \in \{0, 1\}$, one term always contains a $\log(1 - \delta) \approx 0$ and another term always contains a $\log \delta$. The latter term is the problem: it is sent toward negative infinity as the nudging factor δ goes to 0.

As a result, manipulating δ in even minute increments can massively change the likelihood computation so that it pays more or less attention to 0's and 1's—in the limit, ignoring all responses $x_i \in (0, 1)$. Moreover, because the WAIC is computed in terms of the likelihood, manipulating δ can greatly affects model comparison results. Thus introducing such a nudge factor is problematic, given the methodological problems it poses; especially, when there are likelihoods which are actually *defined* for $x_i \in \{0, 1\}$. One such likelihood is given by the ordered beta distribution, which we look at next.

B Ordered beta distributions as likelihoods

Ordered betas models are a variant of the more general 0-1 inflated beta model (see Liu and Eugenio 2018 and references therein) that avoids some of the conceptual problems of the latter (Kubinec 2023). Intuitively, a 0-1 inflated beta likelihood models the response as a two-stage process: first, it decides whether to respond with 0 (which it does with probability $p_{\{0\}}$), with 1 (with probability $p_{\{1\}}$), or with some value in $(0, 1)$ (with probability $p_{(0,1)} = 1 - p_{\{0\}} - p_{\{1\}}$); second, if it decides to respond with some value in $(0, 1)$, it samples that value from a $\mathrm{Beta}(\alpha, \beta)$, where $\alpha \equiv \phi\mu$, $\beta \equiv \phi(1 - \mu)$, $\mu \in (0, 1)$, and $\phi \in \mathbb{R}_+$.

The main challenge of using a standard 0-1 inflated beta likelihood is determining how to define $p_{\{0\}}$ and $p_{\{1\}}$. The easiest way to define these parameters is to allow them to be directly estimated. By allowing this, however, we would effectively be injecting multiple discrete components into the model.

A preferable alternative is to use an ordered beta likelihood. Such a likelihood estimates $p_{\{0\}}$ and $p_{\{1\}}$ in terms of the same μ used in the beta component of the model. The idea is that, when

μ is close to 1, the mean of the beta component will be close to 1, while $p_{\{1\}}$ will be higher; and when μ is close to 0, the mean of the beta component will be close to 0, and $p_{\{0\}}$ will be higher.

To accomplish this, the ordered beta model defines parameters $p_{\{0\}}$, $p_{(0,1)}$, and $p_{\{1\}}$ via the linked logit model (common from ordinal regression). In addition to μ and ϕ, it is parameterized by *cutpoints* $\mathbf{c} = \langle \mathbf{c}_1, \mathbf{c}_2 \rangle$. $p_{\{0\}}$, $p_{\{1\}}$, and $p_{(0,1)}$ are then defined as follows:

$$\begin{aligned}
p_{\{0\}} &\equiv \text{logit}^{-1}(\mathbf{c}_1 - \text{logit}(\mu)) \\
p_{\{1\}} &\equiv \text{logit}^{-1}(\text{logit}(\mu) - \mathbf{c}_2) \\
p_{(0,1)} &\equiv 1 - p_{\{0\}} - p_{\{1\}}
\end{aligned}$$

The probability measure is then given by:

$$\mathbb{P}_{\text{OrdBeta}}\left(x \in (a,b) \mid \mu, \phi, \mathbf{c}\right) = \begin{cases} 1 & \text{if } a < 0, b > 1 \\ p_{\{0\}} + p_{(0,1)} \int_0^b \text{Beta}(x \mid \mu, \phi)\, \text{d}x & \text{if } a < 0, b \leq 1 \\ p_{(0,1)} \int_a^b \text{Beta}(x \mid \mu, \phi)\, \text{d}x & \text{if } a \geq 0, b \leq 1 \\ p_{\{1\}} + p_{(0,1)} \int_a^1 \text{Beta}(x \mid \mu, \phi)\, \text{d}x & \text{if } a \geq 0, b > 1 \\ 0 & \text{otherwise} \end{cases}$$

The log-likelihood is then:

$$\mathcal{L}_{\text{OrdBeta}}(\mu, \phi, \mathbf{c} \mid \mathbf{x}) = \sum_{i:x_i=0} \log p_{\{0\}} + \sum_{i:x_i=1} \log p_{\{1\}} + \sum_{i:x_i \notin \{0,1\}} \log p_{(0,1)} + \log \text{Beta}(x_i \mid \mu, \phi)$$

Here, $\text{Beta}(x_i \mid \mu, \phi)$ is the beta likelihood. Note that this likelihood can be thought of as a conceptually valid variant of the problematic likelihood described in Appendix A.

We provide models fit using this definition of the likelihood in Appendix C.2.

C Full model specifications

C.1 The truncation models

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.

C.1.1 The norming model

We characterize our model of the norming data as a probabilistic program, with the following structure, given data $\boldsymbol{y}_{\text{norming}} : r^{\frac{n_{\text{context}}}{2} * n_{\text{participant}}}$, 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.

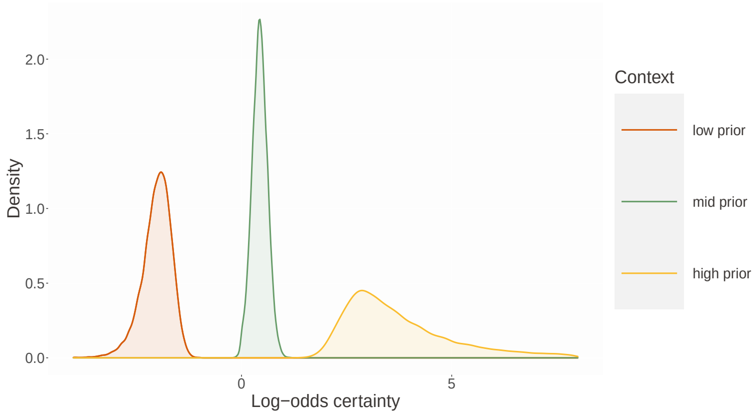


Figure 13: Density plots of the posterior log-odds certainty (with participant intercepts zeroed out) for three items in Degen and Tonhauser's (2021) norming task. Low and high priors are for *Grace visited her sister*, given the facts *Grace hates her sister* and *Grace loves her sister*, respectively. Mid prior is for *Sophia got a tattoo*, given the fact *Sophia is a hipster*.

We encode the certainties for contexts as parameters $\boldsymbol{\omega}$ on a log-odds scale, with participant random intercepts $\boldsymbol{\epsilon}$ added to these parameters before they are mapped to transformed parameters $\boldsymbol{w}$ 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 $\boldsymbol{\sigma_{\omega}}$) of these normal distributions are, in turn, given exponential hyper-priors. 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 $\boldsymbol{w}_{i,j}$ 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

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

to obtain a density (or mass, as the case may be) function on α's from a probabilistic program that returns α's as values. Thus if $\boldsymbol{m}$ returns, say, tuples of real numbers, then we may obtain the density (or mass) that $\boldsymbol{m}$ assigns to the tuple $\boldsymbol{x}$ as $D_{\boldsymbol{m}}(\boldsymbol{x})$.

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

$$\begin{array}{ll}
\texttt{norming-gradient} : \mathsf{P}(r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{context}}} \times r^2) & \\
\texttt{norming-gradient} = & \boldsymbol{\sigma_{\omega}} \sim \texttt{Exponential}(1) \\
& \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1) \\
& \sigma_e \sim \texttt{Uniform}(0, 1) \\
& \boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\omega}}) \\
& \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}}) \\
& \texttt{factor}(D_{\mathcal{N}(\boldsymbol{w}, \sigma_e)\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{norming}})) \\
& \boxed{\langle \boldsymbol{\omega}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\omega}}, \sigma_{\epsilon_{\omega}}, \sigma_e \rangle} \\
& \qquad \texttt{where } \boldsymbol{w}_{i,j} = \texttt{logit}^{-1}(\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_j)
\end{array}$$

The parameters $\boldsymbol{\omega}$ 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 $\boldsymbol{\omega}$), taking a normal distribution with mean and variance equal to that of the posterior distribution. Density plots for three items are given in Figure 13 (see Figure 16 of Appendix D for all items).

For completeness, we also give the discrete model of the norming data. It can be presented compactly as follows:

$$
\begin{aligned}
\texttt{norming-discrete} &: \mathsf{P}(r^{n_{\text{context}}} \times r^{n_{\text{participant}}} \times r^{n_{\text{context}}} \times r^2) \\
\texttt{norming-discrete} &= \; \boldsymbol{\sigma}_{\boldsymbol{\omega}} \sim \texttt{Exponential}(1) \\
&\quad\; \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}} \sim \texttt{Exponential}(1) \\
&\quad\; \sigma_e \sim \texttt{Uniform}(0, 1) \\
&\quad\; \boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{\sigma}_{\boldsymbol{\omega}}) \\
&\quad\; \boldsymbol{\epsilon}_{\boldsymbol{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}) \\
&\quad\; \boldsymbol{\tau}_{\boldsymbol{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\
&\quad\; \texttt{factor}(D_{\mathcal{N}(\mathbb{1}(\boldsymbol{\tau}_{\boldsymbol{w}}), \sigma_e)\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{norming}})) \\
&\quad\; \boxed{\langle \boldsymbol{\omega}, \boldsymbol{\epsilon}_{\boldsymbol{\omega}}, \boldsymbol{\sigma}_{\boldsymbol{\omega}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}, \sigma_e \rangle} \\
&\qquad\quad \texttt{where } \boldsymbol{w}_{i,j} = \texttt{logit}^{-1}(\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_j)
\end{aligned}
$$

C.1.2 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 $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\omega}}$ 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, $\boldsymbol{y}_{\text{projection}} : r^{n_{\text{verb}} * n_{\text{participant}}}$ 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 $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$. Participant random intercepts $\boldsymbol{\epsilon}_{\boldsymbol{\nu}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{\omega}}$ are added to these parameters, respectively, before they are mapped to transformed parameters $\boldsymbol{v}$ and $\boldsymbol{w}$ 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 ($\sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}$, $\sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}$, and $\boldsymbol{\sigma}_{\boldsymbol{\nu}}$) 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 $\boldsymbol{v}_{i,k}$ and $\boldsymbol{w}_{j,k}$ to denote the parameters encoding the probability of projection and certainty, respectively, for participant k, given verb i and context j.

Obtaining a disjunction Recall the claim of Section 4.1 that the following equivalence obtains for the discrete-factivity and wholly-gradient models (see Footnote 17):

$$
\begin{array}{rl}
& m \sim \texttt{commonGround} \\
& x \sim \boxed{\mathbb{P}\left(\begin{array}{l}\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle \sim m \\ \texttt{observe}(\llbracket \textit{Susan knows that Grace visited her sister} \rrbracket^{\langle \boldsymbol{\tau_w},\boldsymbol{\tau_v}\rangle}) \\ \boxed{\llbracket \textit{Grace visited her sister}\rrbracket^{\langle \boldsymbol{\tau_w},\boldsymbol{\tau_v}\rangle}}\end{array}\right)} \\
& f(x,\Psi) \\
= & m \sim \texttt{commonGround} \\
& x \sim \boxed{\mathbb{P}\left(\begin{array}{l}\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle \sim m \\ \boxed{\mathsf{know}(\boldsymbol{\tau_v}) \vee \mathsf{grace}(\boldsymbol{\tau_w})}\end{array}\right)} \\
& f(x,\Psi)
\end{array}
$$

where $\mathsf{know}(\boldsymbol{\tau_v})$ is the component of $\boldsymbol{\tau_v}$ determining whether or not the content of *know* projects and $\mathsf{grace}(\boldsymbol{\tau_w})$ is the component of $\boldsymbol{\tau_w}$ determining whether *Grace visited her sister* is true or false. We can justify this claim as follows.

Note that the first probabilistic program can be rephrased:

$$
\begin{array}{l}
m \sim \texttt{commonGround} \\
x \sim \boxed{\mathbb{P}\left(\begin{array}{l}\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle \sim m \\ \texttt{observe}(\mathsf{know}(\boldsymbol{\tau_v}) \rightarrow \mathsf{grace}(\boldsymbol{\tau_w})) \\ \boxed{\llbracket \textit{Grace visited her sister}\rrbracket^{\langle \boldsymbol{\tau_w},\boldsymbol{\tau_v}\rangle}}\end{array}\right)} \\
f(x,\Psi)
\end{array}
$$

That is, if the content of the complement of *know* projects (i.e., if $\mathsf{know}(\boldsymbol{\tau_v}) = \mathsf{T}$), then according to the semantics assigned to *know*, *Grace visited her sister* is entailed (i.e., $\llbracket \textit{Grace visited her sister}\rrbracket^{\langle \boldsymbol{\tau_w},\boldsymbol{\tau_v}\rangle} = \mathsf{grace}(\boldsymbol{\tau_w}) = \mathsf{T}$).

Indeed, due to the latter equivalence, the probabilistic program may be rephrased again:

$$
\begin{array}{l}
m \sim \texttt{commonGround} \\
x \sim \boxed{\mathbb{P}\left(\begin{array}{l}\langle \boldsymbol{\tau_w}, \boldsymbol{\tau_v}\rangle \sim m \\ \texttt{observe}(\mathsf{know}(\boldsymbol{\tau_v}) \rightarrow \mathsf{grace}(\boldsymbol{\tau_w})) \\ \boxed{\mathsf{grace}(\boldsymbol{\tau_w})}\end{array}\right)} \\
f(x,\Psi)
\end{array}
$$

From this step, we can derive a disjunction by doing case analysis on the individual models.

The discrete-factivity model In this case, the above probabilistic program may be specified as follows:

$$
\begin{array}{l}
\langle \boldsymbol{w}, \boldsymbol{\tau_v}\rangle \sim \texttt{commonGround} \\
x \sim \boxed{\mathbb{P}\left(\begin{array}{l}\boldsymbol{\tau_w} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\ \texttt{observe}(\mathsf{know}(\boldsymbol{\tau_v}) \rightarrow \mathsf{grace}(\boldsymbol{\tau_w})) \\ \boxed{\mathsf{grace}(\boldsymbol{\tau_w})}\end{array}\right)} \\
f(x,\Psi)
\end{array}
$$

When $\boldsymbol{\tau_v}$ is in scope, $\mathsf{know}(\boldsymbol{\tau_v})$ is either T or F. When $\mathsf{know}(\boldsymbol{\tau_v}) = \mathsf{F}$, the implication inside the observe statement is true and may be removed; meanwhile, $\boxed{\mathsf{grace}(\boldsymbol{\tau_w})}$ becomes equivalent to $\boxed{\mathsf{know}(\boldsymbol{\tau_v}) \vee \mathsf{grace}(\boldsymbol{\tau_w})}$. When $\mathsf{know}(\boldsymbol{\tau_v}) = \mathsf{T}$, the formula inside the observe statement is equivalent to $\mathsf{grace}(\boldsymbol{\tau_w})$. As a result, the probability computed is 1, so that it is equivalent to taking the probability of a disjunction one of whose disjuncts is $\mathsf{T} = \mathsf{know}(\boldsymbol{\tau_v})$.

From here, we can determine that 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.

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

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:

$$\begin{array}{ll}
\mathtt{discrete\text{-}factivity} : \mathsf{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{2n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{3}) & \\
\mathtt{discrete\text{-}factivity} = & \boldsymbol{\sigma_\nu} \sim \mathtt{Exponential}(1) \\
& \sigma_{\epsilon_\nu} \sim \mathtt{Exponential}(1) \\
& \sigma_{\epsilon_\omega} \sim \mathtt{Exponential}(1) \\
& \sigma_e \sim \mathtt{Uniform}(0,1) \\
& \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_\nu}) \\
& \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_\omega}, \boldsymbol{\sigma_\omega}) \\
& \boldsymbol{\epsilon_\nu} \sim \mathcal{N}(0, \sigma_{\epsilon_\nu}) \\
& \boldsymbol{\epsilon_\omega} \sim \mathcal{N}(0, \sigma_{\epsilon_\omega}) \\
& \boldsymbol{\tau_v} \sim \mathtt{Bernoulli}(\boldsymbol{v}) \\
& \mathtt{factor}(D_{\mathcal{N}(\boldsymbol{\theta},\sigma_e)\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}})) \\
& \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_\nu}, \boldsymbol{\epsilon_\omega}, \boldsymbol{\sigma_\nu}, \sigma_{\epsilon_\nu}, \sigma_{\epsilon_\omega}, \sigma_e \rangle} \\
& \quad \mathtt{where} \;\; \begin{array}{l} \boldsymbol{v}_{i,k} = \mathtt{logit}^{-1}(\boldsymbol{\nu}_i + \boldsymbol{\epsilon_\nu}_k) \\ \boldsymbol{w}_{j,k} = \mathtt{logit}^{-1}(\boldsymbol{\omega}_j + \boldsymbol{\epsilon_\omega}_k) \\ \boldsymbol{\theta}_{i,j,k} = \mathbb{1}(\boldsymbol{\tau_v}_{i,k}) + \mathbb{1}(\neg\boldsymbol{\tau_v}_{i,k}) * \boldsymbol{w}_{j,k} \end{array}
\end{array}$$

Note that adding an anti-veridicality component for each predicate to the discrete-factivity model involves the following adjustments:

$$
\begin{array}{ll}
\texttt{discrete-factivity+a-v} : \mathsf{P}(r^{2n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{3n_{\text{participant}}} \times r^{2n_{\text{verb}}} \times r^{4}) & \\
\texttt{discrete-factivity+a-v} = & \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
& \boldsymbol{\sigma_{\alpha}} \sim \texttt{Exponential}(1) \\
& \sigma_{\epsilon_{\nu}} \sim \texttt{Exponential}(1) \\
& \sigma_{\epsilon_{\alpha}} \sim \texttt{Exponential}(1) \\
& \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1) \\
& \sigma_{e} \sim \texttt{Uniform}(0,1) \\
& \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
& \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\alpha}}) \\
& \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
& \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}}) \\
& \boldsymbol{\epsilon_{\alpha}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}}) \\
& \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}}) \\
& \boldsymbol{\tau_{v}} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\
& \boldsymbol{\tau_{a}} \sim \texttt{Bernoulli}(\boldsymbol{a}) \\
& \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \sigma_e)\top[0,1]}(\boldsymbol{y}_{\text{projection}})) \\
& \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\alpha}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \boldsymbol{\sigma_{\alpha}}, \sigma_{\epsilon_{\nu}}, \sigma_{\epsilon_{\alpha}}, \sigma_{\epsilon_{\omega}}, \sigma_{e} \rangle} \\
& \qquad \texttt{where} \quad \begin{aligned}[t] \boldsymbol{v}_{i,k} &= \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon_{\nu}}_{k}) \\ \boldsymbol{a}_{i,k} &= \texttt{logit}^{-1}(\boldsymbol{\alpha}_{i} + \boldsymbol{\epsilon_{\alpha}}_{k}) \\ \boldsymbol{w}_{j,k} &= \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon_{\omega}}_{k}) \\ \boldsymbol{\theta}_{i,j,k} &= \mathbb{1}(\boldsymbol{\tau_{v}}_{i,k}) + \mathbb{1}(\neg \boldsymbol{\tau_{v}}_{i,k} \wedge \neg \boldsymbol{\tau_{a}}_{i,k}) * \boldsymbol{w}_{j,k} \end{aligned}
\end{array}
$$

The wholly-gradient model In this case, the probabilistic program of Appendix C.1.2 can be specified as follows:

$$
\begin{array}{l}
\langle \boldsymbol{w}, \boldsymbol{v} \rangle \sim \texttt{commonGround} \\
x \sim \boxed{\mathbb{P}\left(\begin{array}{l} \boldsymbol{\tau_{v}} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\ \boldsymbol{\tau_{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\ \mathsf{observe}(\mathsf{know}(\boldsymbol{\tau_{v}}) \rightarrow \mathsf{grace}(\boldsymbol{\tau_{w}})) \\ \boxed{\mathsf{grace}(\boldsymbol{\tau_{w}})} \end{array} \right)} \\
f(x, \Psi)
\end{array}
$$

The reasoning here is similar to that for the discrete-factivity model. The crucial difference is that $\boldsymbol{\tau_{v}}$ is now only in scope after the introduction of of the probability operator. But the two cases to consider ($\mathsf{know}(\boldsymbol{\tau_{v}}) = \mathsf{T}$ and $\mathsf{know}(\boldsymbol{\tau_{v}}) = \mathsf{F}$) are analogous.

From here, we can determine that the wholly-gradient model sets each parameter $\boldsymbol{\theta}_{i,j,k}$ equal to $\boldsymbol{v}_{i,k} + (1 - \boldsymbol{v}_{i,k}) * \boldsymbol{w}_{j,k}$, an encoding justified by the following fact:

Fact 2.

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

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).

$$
\begin{aligned}
&\texttt{wholly-gradient} : \mathsf{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{2n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{3})\\
&\texttt{wholly-gradient} = \; \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{\epsilon_{\nu}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{e} \sim \texttt{Uniform}(0,1)\\
&\qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}})\\
&\qquad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}})\\
&\qquad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}})\\
&\qquad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}})\\
&\qquad \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta},\sigma_{e})\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}}))\\
&\qquad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \sigma_{\epsilon_{\nu}}, \sigma_{\epsilon_{\omega}}, \sigma_{e} \rangle}\\
&\qquad\quad \texttt{where} \quad \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon_{\nu}}_{k})\\
&\qquad\qquad\qquad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon_{\omega}}_{k})\\
&\qquad\qquad\qquad \boldsymbol{\theta}_{i,j,k} = \boldsymbol{v}_{i,k} + (1 - \boldsymbol{v}_{i,k}) * \boldsymbol{w}_{j,k}
\end{aligned}
$$

Note that adding an anti-veridicality component for each predicate to the wholly-gradient model involves the following adjustments:

$$
\begin{aligned}
&\texttt{wholly-gradient+a-v} : \mathsf{P}(r^{2n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{3n_{\text{participant}}} \times r^{2n_{\text{verb}}} \times r^{4})\\
&\texttt{wholly-gradient+a-v} = \; \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1)\\
&\qquad \boldsymbol{\sigma_{\alpha}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{\epsilon_{\nu}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{\epsilon_{\alpha}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1)\\
&\qquad \sigma_{e} \sim \texttt{Uniform}(0,1)\\
&\qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}})\\
&\qquad \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\alpha}})\\
&\qquad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}})\\
&\qquad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}})\\
&\qquad \boldsymbol{\epsilon_{\alpha}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}})\\
&\qquad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}})\\
&\qquad \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta},\sigma_{e})\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}}))\\
&\qquad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\alpha}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \boldsymbol{\sigma_{\alpha}}, \sigma_{\epsilon_{\nu}}, \sigma_{\epsilon_{\alpha}}, \sigma_{\epsilon_{\omega}}, \sigma_{e} \rangle}\\
&\qquad\quad \texttt{where} \quad \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon_{\nu}}_{k})\\
&\qquad\qquad\qquad \boldsymbol{a}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\alpha}_{i} + \boldsymbol{\epsilon_{\alpha}}_{k})\\
&\qquad\qquad\qquad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon_{\omega}}_{k})\\
&\qquad\qquad\qquad \boldsymbol{\theta}_{i,j,k} = \boldsymbol{v}_{i,k} + (1 - \boldsymbol{v}_{i,k}) * (1 - \boldsymbol{a}_{i,k}) * \boldsymbol{w}_{j,k}
\end{aligned}
$$

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.

$$
\begin{aligned}
&\texttt{discrete-world} : \mathsf{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{2n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{3}) \\
&\texttt{discrete-world} = \; \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{e} \sim \texttt{Uniform}(0, 1) \\
&\qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
&\qquad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
&\qquad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}) \\
&\qquad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}) \\
&\qquad \boldsymbol{\tau_{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\
&\qquad \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \sigma_{e})\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}})) \\
&\qquad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}, \sigma_{e} \rangle} \\
&\qquad\qquad \texttt{where} \;\; \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon}_{\boldsymbol{\nu}k}) \\
&\qquad\qquad\qquad\quad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon}_{\boldsymbol{\omega}k}) \\
&\qquad\qquad\qquad\quad \boldsymbol{\theta}_{i,j,k} = \mathbb{1}(\boldsymbol{\tau}_{\boldsymbol{w}\,j,k}) + \mathbb{1}(\neg\boldsymbol{\tau}_{\boldsymbol{w}\,j,k}) * \boldsymbol{v}_{i,k}
\end{aligned}
$$

Note that adding an anti-veridicality component for each predicate to the discrete-world model does not change it.

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

$$
\begin{aligned}
&\texttt{wholly-discrete} : \mathsf{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{2n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{3}) \\
&\texttt{wholly-discrete} = \; \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{e} \sim \texttt{Uniform}(0, 1) \\
&\qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
&\qquad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
&\qquad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}) \\
&\qquad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}) \\
&\qquad \boldsymbol{\tau_{v}} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\
&\qquad \boldsymbol{\tau_{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\
&\qquad \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \sigma_{e})\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}})) \\
&\qquad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}, \sigma_{e} \rangle} \\
&\qquad\qquad \texttt{where} \;\; \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon}_{\boldsymbol{\nu}k}) \\
&\qquad\qquad\qquad\quad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon}_{\boldsymbol{\omega}k}) \\
&\qquad\qquad\qquad\quad \boldsymbol{\theta}_{i,j,k} = \mathbb{1}(\boldsymbol{\tau}_{\boldsymbol{v}\,i,k} \vee \boldsymbol{\tau}_{\boldsymbol{w}\,j,k})
\end{aligned}
$$

Note that adding an anti-veridicality component for each predicate to the wholly-discrete model involves the following adjustments:

$$
\begin{aligned}
\texttt{wholly-discrete+a-v} &: \mathsf{P}(r^{2n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{3n_{\text{participant}}} \times r^{2n_{\text{verb}}} \times r^{4}) \\
\texttt{wholly-discrete+a-v} &= \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
&\quad \boldsymbol{\sigma_{\alpha}} \sim \texttt{Exponential}(1) \\
&\quad \sigma_{\epsilon_{\nu}} \sim \texttt{Exponential}(1) \\
&\quad \sigma_{\epsilon_{\alpha}} \sim \texttt{Exponential}(1) \\
&\quad \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1) \\
&\quad \sigma_{e} \sim \texttt{Uniform}(0, 1) \\
&\quad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
&\quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\alpha}}) \\
&\quad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
&\quad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}}) \\
&\quad \boldsymbol{\epsilon_{\alpha}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}}) \\
&\quad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}}) \\
&\quad \boldsymbol{\tau_{v}} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\
&\quad \boldsymbol{\tau_{a}} \sim \texttt{Bernoulli}(\boldsymbol{a}) \\
&\quad \boldsymbol{\tau_{w}} \sim \texttt{Bernoulli}(\boldsymbol{w}) \\
&\quad \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta}, \sigma_e)\mathsf{T}[0,1]}(\boldsymbol{y}_{\text{projection}})) \\
&\quad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\alpha}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\sigma_{\nu}}, \boldsymbol{\sigma_{\alpha}}, \sigma_{\epsilon_{\nu}}, \sigma_{\epsilon_{\alpha}}, \sigma_{\epsilon_{\omega}}, \sigma_e \rangle} \\
&\qquad\qquad \texttt{where} \quad \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_i + \boldsymbol{\epsilon_{\nu}}_k) \\
&\qquad\qquad\qquad\quad \boldsymbol{a}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\alpha}_i + \boldsymbol{\epsilon_{\alpha}}_k) \\
&\qquad\qquad\qquad\quad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_j + \boldsymbol{\epsilon_{\omega}}_k) \\
&\qquad\qquad\qquad\quad \boldsymbol{\theta}_{i,j,k} = \mathbb{1}(\boldsymbol{\tau_{v}}_{i,k} \vee (\neg \boldsymbol{\tau_{a}}_{i,k} \wedge \boldsymbol{\tau_{w}}_{j,k}))
\end{aligned}
$$

C.1.3 The contentful evaluation models

To evaluate the four models using this data, we obtained, from each model, means $\boldsymbol{\mu_{\nu}}$ and standard deviations $\boldsymbol{\sigma_{\nu}}$ of the marginal posterior log-odds of projection distributions for predicates, as well as means $\boldsymbol{\mu_{\omega}}$ and standard deviations $\boldsymbol{\sigma_{\omega}}$ 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 14 for density plots of these posterior distributions for six predicates, and Figure 17 of Appendix D for density plots of the posterior disributions for all predicates.)

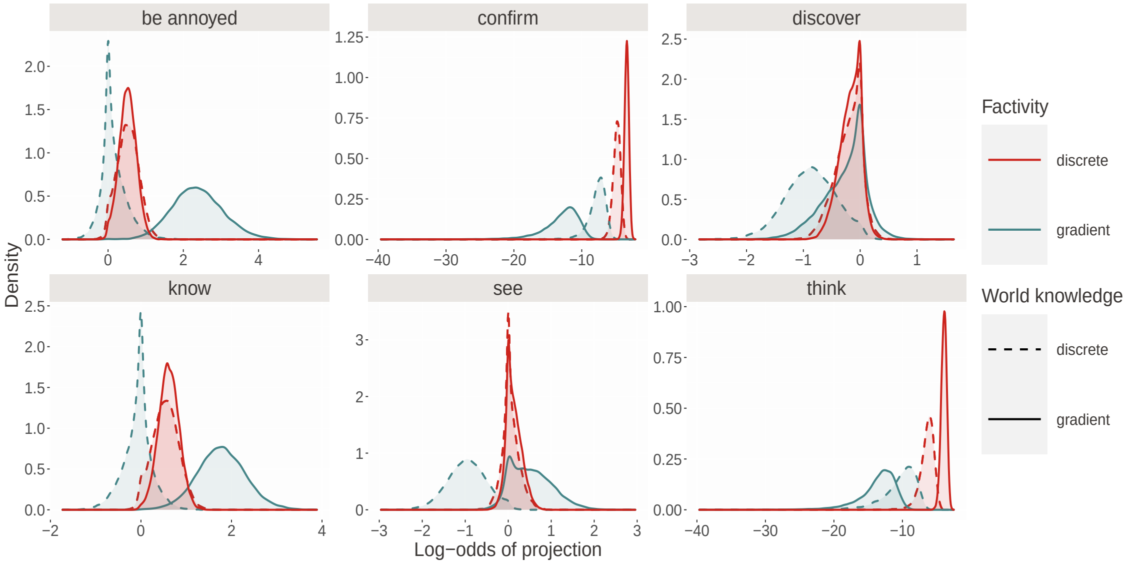


Figure 14: Density plots of the posterior log-odds of projection (with participant intercepts zeroed out) for all four models for six predicates from Degen and Tonhauser's (2021) projection experiment.

Each evaluation has the following structure:

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

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

C.1.4 The non-contentful evaluation models

To evaluate the four models using both the bleached and templatic data, we used the means $\boldsymbol{\mu}_\nu$ and standard deviations $\boldsymbol{\sigma}_\nu$ 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 structure, where ellipses, as above, indicate the unique aspects of each of the four models:

$$
\begin{aligned}
\texttt{non-contentful-evaluation} &: \mathsf{P}(r^{n_{\text{verb}}} \times r \times r^{2n_{\text{participant}}} \times r^{4})\\
\texttt{non-contentful-evaluation} &= \sigma_\omega \sim \texttt{Exponential}(1)\\
&\ \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}} \sim \texttt{Exponential}(1)\\
&\ \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}} \sim \texttt{Exponential}(1)\\
&\ \sigma_e \sim \texttt{Uniform}(0,1)\\
&\ \boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\nu}}, \boldsymbol{\sigma}_{\boldsymbol{\nu}})\\
&\ \omega \sim \mathcal{N}(0, \sigma_\omega)\\
&\ \boldsymbol{\epsilon}_{\boldsymbol{\nu}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}})\\
&\ \boldsymbol{\epsilon}_{\boldsymbol{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}})\\
&\ \vdots\\
&\ \texttt{factor}(D_{\mathcal{N}(\boldsymbol{\theta},\sigma_e)\text{T}[0,1]}(\boldsymbol{y}_{\text{non-contentful}}))\\
&\ \boxed{\langle \boldsymbol{\nu}, \omega, \boldsymbol{\epsilon}_{\boldsymbol{\nu}}, \boldsymbol{\epsilon}_{\boldsymbol{\omega}}, \sigma_\omega, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\nu}}}, \sigma_{\boldsymbol{\epsilon}_{\boldsymbol{\omega}}}, \sigma_e \rangle}\\
&\ \qquad \texttt{where}\quad \boldsymbol{v}_{i,j} = \texttt{logit}^{-1}(\boldsymbol{\nu}_i + \boldsymbol{\epsilon}_{\boldsymbol{\nu}\,j})\\
&\ \qquad \phantom{\texttt{where}}\quad \boldsymbol{w}_j = \texttt{logit}^{-1}(\omega + \boldsymbol{\epsilon}_{\boldsymbol{\omega}\,j})\\
&\ \qquad \phantom{\texttt{where}}\quad \boldsymbol{\theta}_{i,j} = \ldots
\end{aligned}
$$

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

C.2 The inflation models

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 18,000 posterior samples of the model parameters, following 18,000 burn-in samples, on four chains each.

In general, the inflation models are identical to their truncation counterparts, the only differences pertaining to the definition of the likelihood. Note that we obtained reliable fits for the norming model, as well as for the discrete-factivity model and the wholly-gradient model; we obtained unreliable fits for the discrete-world and wholly-discrete models, which suffered from poor convergence diagnostics, despite sustained effort. We still present the latter two, anyway, for completeness. Crucially, we obtained reliable results for the two models which enter into our main comparison.

C.2.1 The norming model

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

$$\texttt{norming} : \mathsf{P}(r^{n_{\text{context}}} \times r^{3n_{\text{participant}}} \times r^{n_{\text{context}}} \times r^{5})$$

$$\begin{aligned}
\texttt{norming} = \; & \boldsymbol{\sigma_{\omega}} \sim \texttt{Exponential}(1) \\
& \sigma_{\boldsymbol{\epsilon_{\omega}}} \sim \texttt{Exponential}(1) \\
& \sigma_{\boldsymbol{\epsilon_{\kappa,1}}} \sim \texttt{Exponential}(1) \\
& \sigma_{\boldsymbol{\epsilon_{\kappa,2}}} \sim \texttt{Exponential}(1) \\
& \kappa \sim \mathcal{N}(\texttt{log}(4), 1) \\
& \phi \sim \texttt{Exponential}(0.1) \\
& \boldsymbol{\omega} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\omega}}) \\
& \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\omega}}}) \\
& \boldsymbol{\epsilon_{\kappa,1}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\kappa,1}}}) \\
& \boldsymbol{\epsilon_{\kappa,2}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\kappa,2}}}) \\
& \texttt{factor}(D_{\texttt{OrdBeta}(\boldsymbol{w},\phi,\mathbf{c})}(\boldsymbol{y}_{\text{norming}})) \\
& \boxed{\langle \boldsymbol{\omega}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\epsilon_{\kappa,1}}, \boldsymbol{\epsilon_{\kappa,2}}, \boldsymbol{\sigma_{\omega}}, \sigma_{\boldsymbol{\epsilon_{\omega}}}, \sigma_{\boldsymbol{\epsilon_{\kappa,1}}}, \sigma_{\boldsymbol{\epsilon_{\kappa,2}}}, \kappa, \phi \rangle} \\
& \qquad \texttt{where} \quad \boldsymbol{w}_{i,j} = \texttt{logit}^{-1}(\boldsymbol{\omega}_i + \boldsymbol{\epsilon_{\omega}}_j) \\
& \qquad \qquad \quad \mathbf{c}_{1,i,j} = -e^{\kappa + \boldsymbol{\epsilon_{\kappa,1}}_j} \\
& \qquad \qquad \quad \mathbf{c}_{2,i,j} = e^{\kappa + \boldsymbol{\epsilon_{\kappa,2}}_j}
\end{aligned}$$

The parameters $\boldsymbol{\omega}$ encode a log-odds certainty rating for each item. Following our strategy for the truncation models, 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 $\boldsymbol{\omega}$), taking a normal distribution with mean and variance equal to that of the posterior distribution.

C.2.2 The factivity models

In our specifications of the factivity models, $\boldsymbol{y}_{\text{projection}} : r^{n_{\text{verb}} * n_{\text{participant}}}$ encodes the experimental data, following our presentation of the truncation models.

The discrete-factivity model

$$
\begin{aligned}
\texttt{discrete-factivity} &: \mathsf{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{4n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{7}) \\
\texttt{discrete-factivity} &= \begin{aligned}[t]
&\boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
&\sigma_{\boldsymbol{\epsilon_{\nu}}} \sim \texttt{Exponential}(1) \\
&\sigma_{\boldsymbol{\epsilon_{\omega}}} \sim \texttt{Exponential}(1) \\
&\sigma_{\boldsymbol{\epsilon_{\kappa}},1} \sim \texttt{Exponential}(1) \\
&\sigma_{\boldsymbol{\epsilon_{\kappa}},2} \sim \texttt{Exponential}(1) \\
&\kappa \sim \mathcal{N}(\texttt{log}(4), 1) \\
&\phi \sim \texttt{Exponential}(0.1) \\
&\eta \sim \mathcal{N}(0, 1) \\
&\boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
&\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
&\boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\nu}}}) \\
&\boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\omega}}}) \\
&\boldsymbol{\epsilon_{\kappa}},_{1} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\kappa}},1}) \\
&\boldsymbol{\epsilon_{\kappa}},_{2} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon_{\kappa}},2}) \\
&\boldsymbol{\tau_{v}} \sim \texttt{Bernoulli}(\boldsymbol{v}) \\
&\texttt{factor}(D_{\texttt{OrdBeta}(\boldsymbol{\theta},\phi,\mathbf{c})}(\boldsymbol{y}_{\text{projection}})) \\
&\boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\epsilon_{\kappa}},_{1}, \boldsymbol{\epsilon_{\kappa}},_{2}, \boldsymbol{\sigma_{\nu}}, \sigma_{\boldsymbol{\epsilon_{\nu}}}, \sigma_{\boldsymbol{\epsilon_{\omega}}}, \sigma_{\boldsymbol{\epsilon_{\kappa}},1}, \sigma_{\boldsymbol{\epsilon_{\kappa}},2}, \kappa, \phi, \eta \rangle} \\
&\qquad \texttt{where} \quad \begin{aligned}[t]
\boldsymbol{v}_{i,k} &= \texttt{logit}^{-1}(\boldsymbol{\nu}_{i} + \boldsymbol{\epsilon_{\nu}}_{k}) \\
\boldsymbol{w}_{j,k} &= \texttt{logit}^{-1}(\boldsymbol{\omega}_{j} + \boldsymbol{\epsilon_{\omega}}_{k}) \\
\mathbf{c}_{1,i,j,k} &= -e^{\kappa + \boldsymbol{\epsilon_{\kappa}},_{1k}} \\
\mathbf{c}_{2,i,j,k} &= e^{\kappa + \boldsymbol{\epsilon_{\kappa}},_{2k}} \\
\mu_{1,i,j,k} &= \texttt{logit}^{-1}(\mathbf{c}_{2,i,j,k} + \eta) \\
\boldsymbol{\theta}_{i,j,k} &= \mathbb{1}(\boldsymbol{\tau_{v}}_{i,k}) * \mu_{1,i,j,k} + \mathbb{1}(\neg \boldsymbol{\tau_{v}}_{i,k}) * \boldsymbol{w}_{j,k}
\end{aligned}
\end{aligned}
\end{aligned}
$$

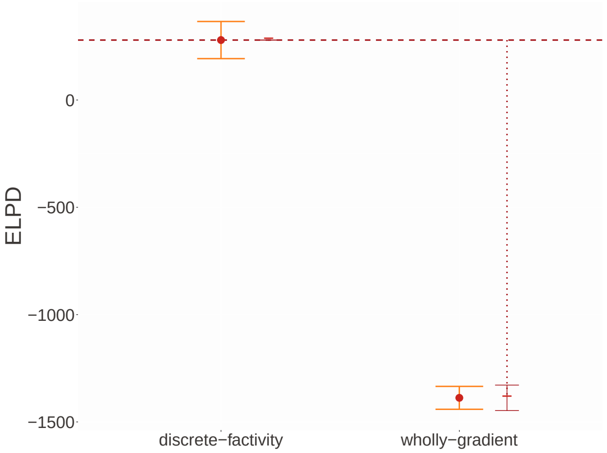


Figure 15: ELPDs for the discrete-factivity and wholly-gradient ordered beta models. Dotted lines indicate estimated differences between each model and the discrete-factivity model. Error bars indicate standard errors.

The wholly-gradient model

$$
\begin{aligned}
&\texttt{wholly-gradient} : \mathtt{P}(r^{n_{\text{verb}}} \times r^{n_{\text{context}}} \times r^{4n_{\text{participant}}} \times r^{n_{\text{verb}}} \times r^{6}) \\
&\texttt{wholly-gradient} = \; \boldsymbol{\sigma_{\nu}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\epsilon_{\nu}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\epsilon_{\omega}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\epsilon_{\kappa,1}} \sim \texttt{Exponential}(1) \\
&\qquad \sigma_{\epsilon_{\kappa,2}} \sim \texttt{Exponential}(1) \\
&\qquad \kappa \sim \mathcal{N}(\texttt{log}(4), 1) \\
&\qquad \phi \sim \texttt{Exponential}(0.1) \\
&\qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\sigma_{\nu}}) \\
&\qquad \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu_{\omega}}, \boldsymbol{\sigma_{\omega}}) \\
&\qquad \boldsymbol{\epsilon_{\nu}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\nu}}) \\
&\qquad \boldsymbol{\epsilon_{\omega}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\omega}}) \\
&\qquad \boldsymbol{\epsilon_{\kappa,1}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\kappa,1}}) \\
&\qquad \boldsymbol{\epsilon_{\kappa,2}} \sim \mathcal{N}(0, \sigma_{\epsilon_{\kappa,2}}) \\
&\qquad \texttt{factor}(D_{\texttt{OrdBeta}(\boldsymbol{\theta},\phi,\mathbf{c})}(\boldsymbol{y}_{\text{projection}})) \\
&\qquad \boxed{\langle \boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{\epsilon_{\nu}}, \boldsymbol{\epsilon_{\omega}}, \boldsymbol{\epsilon_{\kappa,1}}, \boldsymbol{\epsilon_{\kappa,2}}, \boldsymbol{\sigma_{\nu}}, \sigma_{\epsilon_{\nu}}, \sigma_{\epsilon_{\omega}}, \sigma_{\epsilon_{\kappa,1}}, \sigma_{\epsilon_{\kappa,2}}, \kappa, \phi \rangle} \\
&\qquad\qquad \texttt{where} \quad \boldsymbol{v}_{i,k} = \texttt{logit}^{-1}(\boldsymbol{\nu}_i + \boldsymbol{\epsilon_{\nu}}_k) \\
&\qquad\qquad\qquad\quad \boldsymbol{w}_{j,k} = \texttt{logit}^{-1}(\boldsymbol{\omega}_j + \boldsymbol{\epsilon_{\omega}}_k) \\
&\qquad\qquad\qquad\quad \mathbf{c}_{1,i,j,k} = -e^{\kappa + \boldsymbol{\epsilon_{\kappa,1}}_k} \\
&\qquad\qquad\qquad\quad \mathbf{c}_{2,i,j,k} = e^{\kappa + \boldsymbol{\epsilon_{\kappa,2}}_k} \\
&\qquad\qquad\qquad\quad \boldsymbol{\theta}_{i,j,k} = \boldsymbol{v}_{i,k} + (1 - \boldsymbol{v}_{i,k}) * \boldsymbol{w}_{j,k}
\end{aligned}
$$

C.2.3 Comparisons

Figure 15 presents ELPDs for the four models. Again, the discrete model performs substantially better than the gradient model. We can therefore conclude that our initial comparison, which used truncated normal distributions, is unlikely to have favored the discrete-factivity model due to the particular likelihood we chose.

D Plots

D.1 Posterior parameter distributions

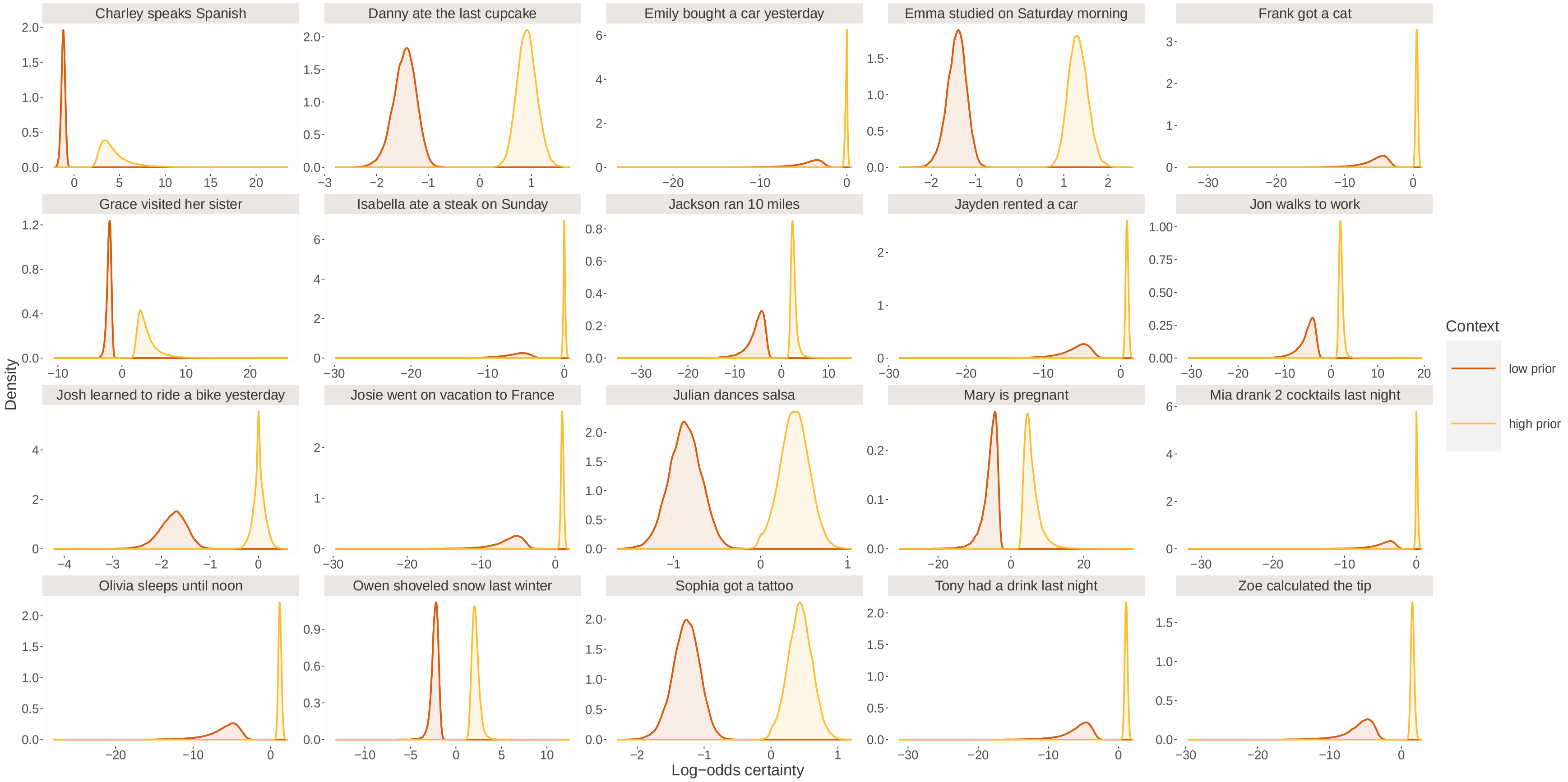


Figure 16: Density plots of the posterior log-odds certainty (with participant intercepts zeroed out) for all items in Degen and Tonhauser's (2021) norming task.

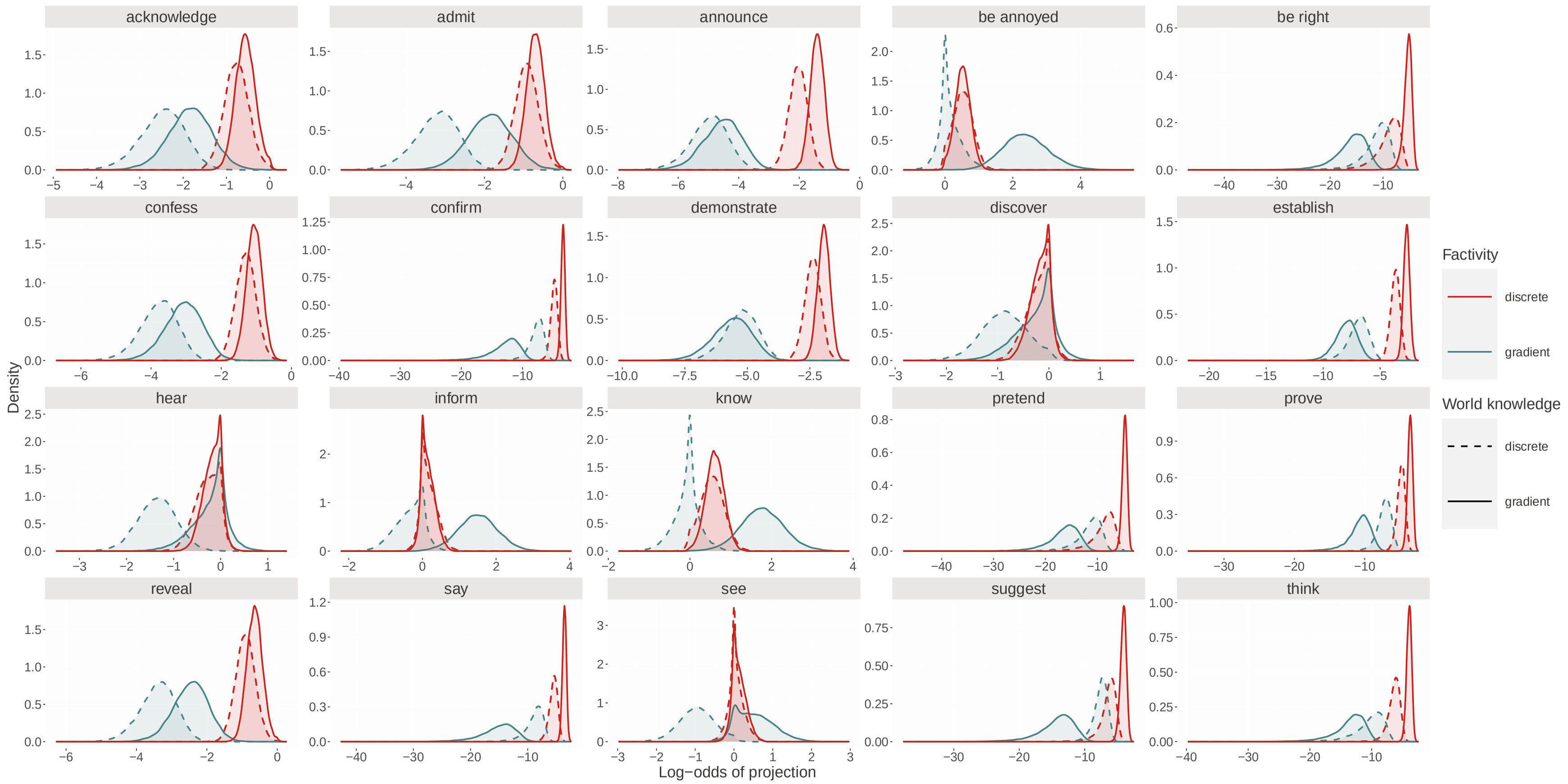


Figure 17: Density plots of the posterior log-odds of projection (with participant intercepts zeroed out) for all four models for all predicates in Degen and Tonhauser's (2021) projection experiment.

D.2 Posterior predictive distributions

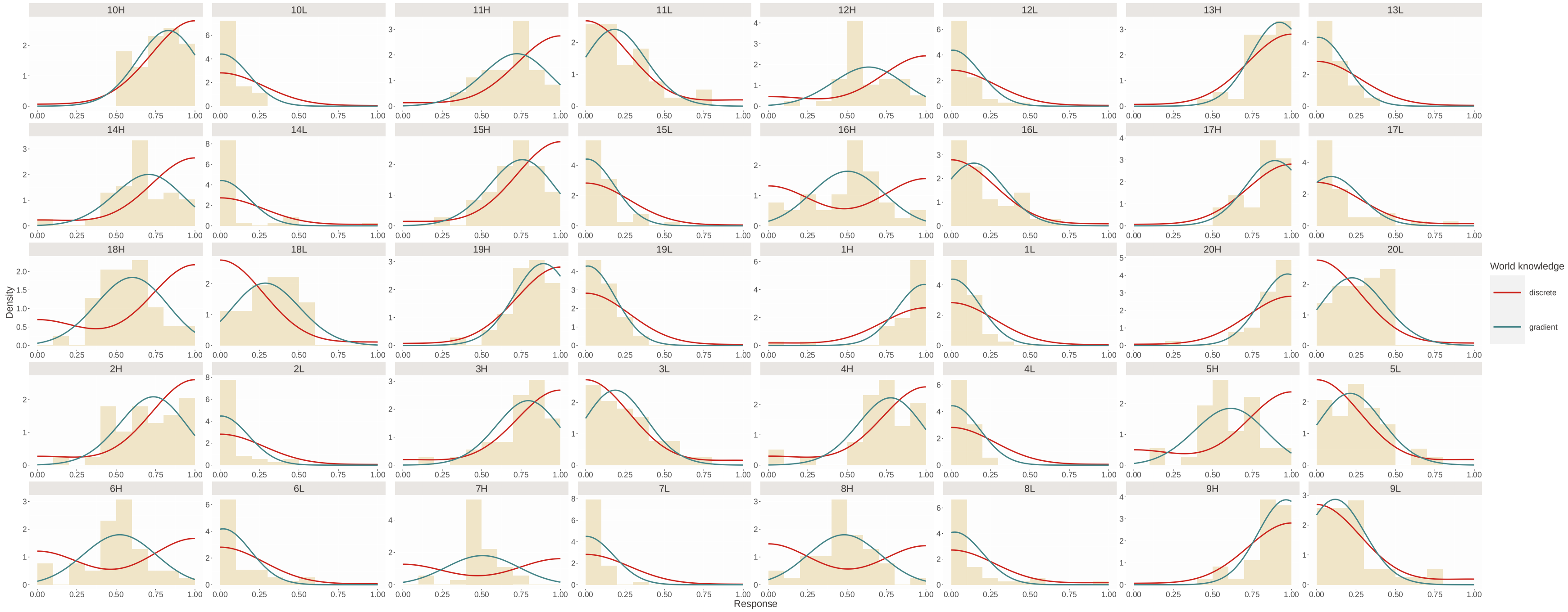


Figure 18: Posterior predictive distributions (with simulated participant intercepts) of the norming-gradient and norming-discrete models for all items in Degen and Tonhauser's (2021) norming experiment. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

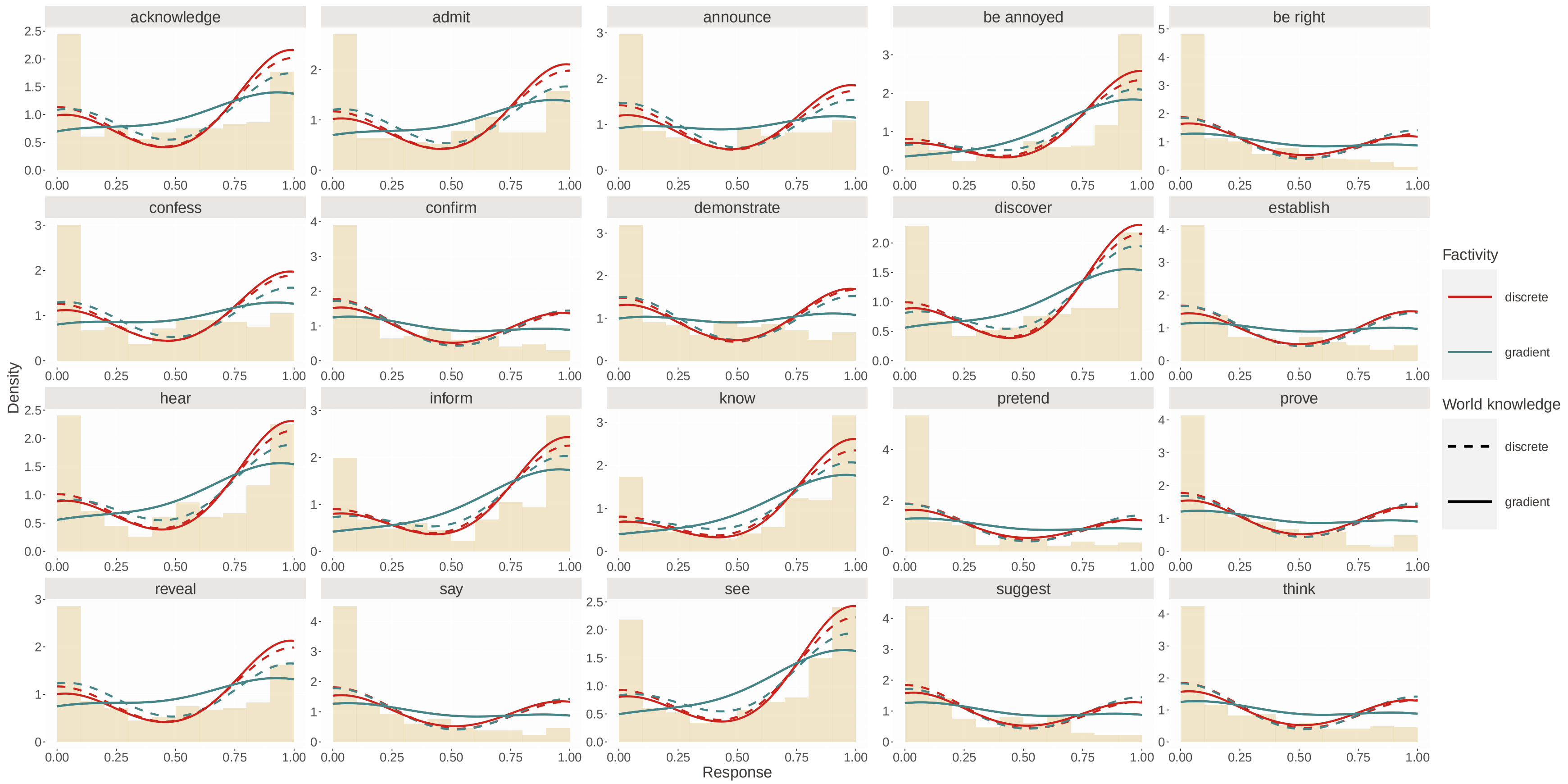


Figure 19: Posterior predictive distributions (with simulated participant intercepts) of all four models for all predicates in Degen and Tonhauser's (2021) projection experiment, for all contexts combined. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

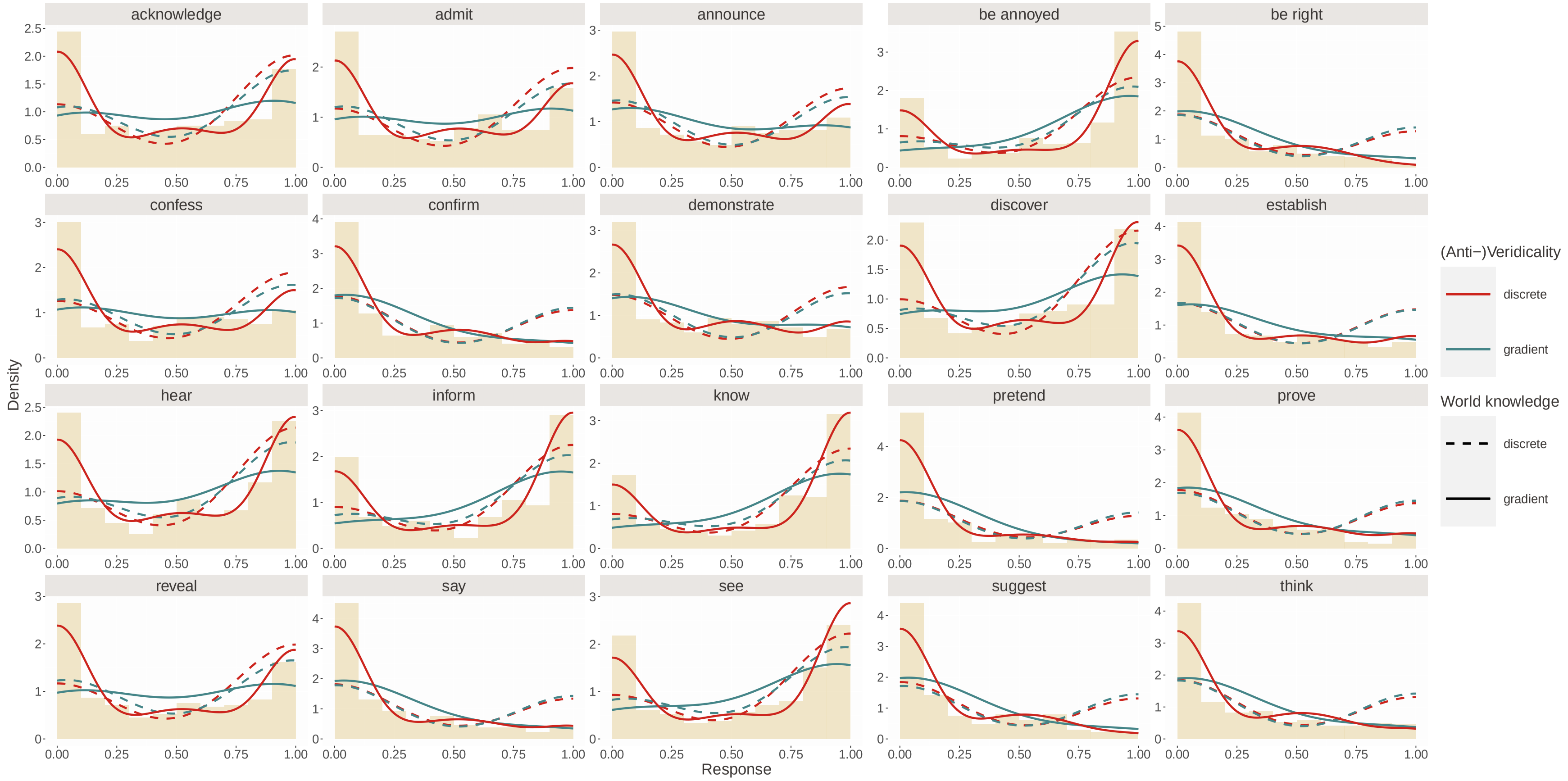


Figure 20: Posterior predictive distributions (with simulated participant intercepts) of all four models for all predicates in Degen and Tonhauser's (2021) projection experiment, for all contexts combined. Here each predicate may also take on an anti-veridical interpretation. Empirical distributions are represented by density histograms of data from Degen and Tonhauser 2021.

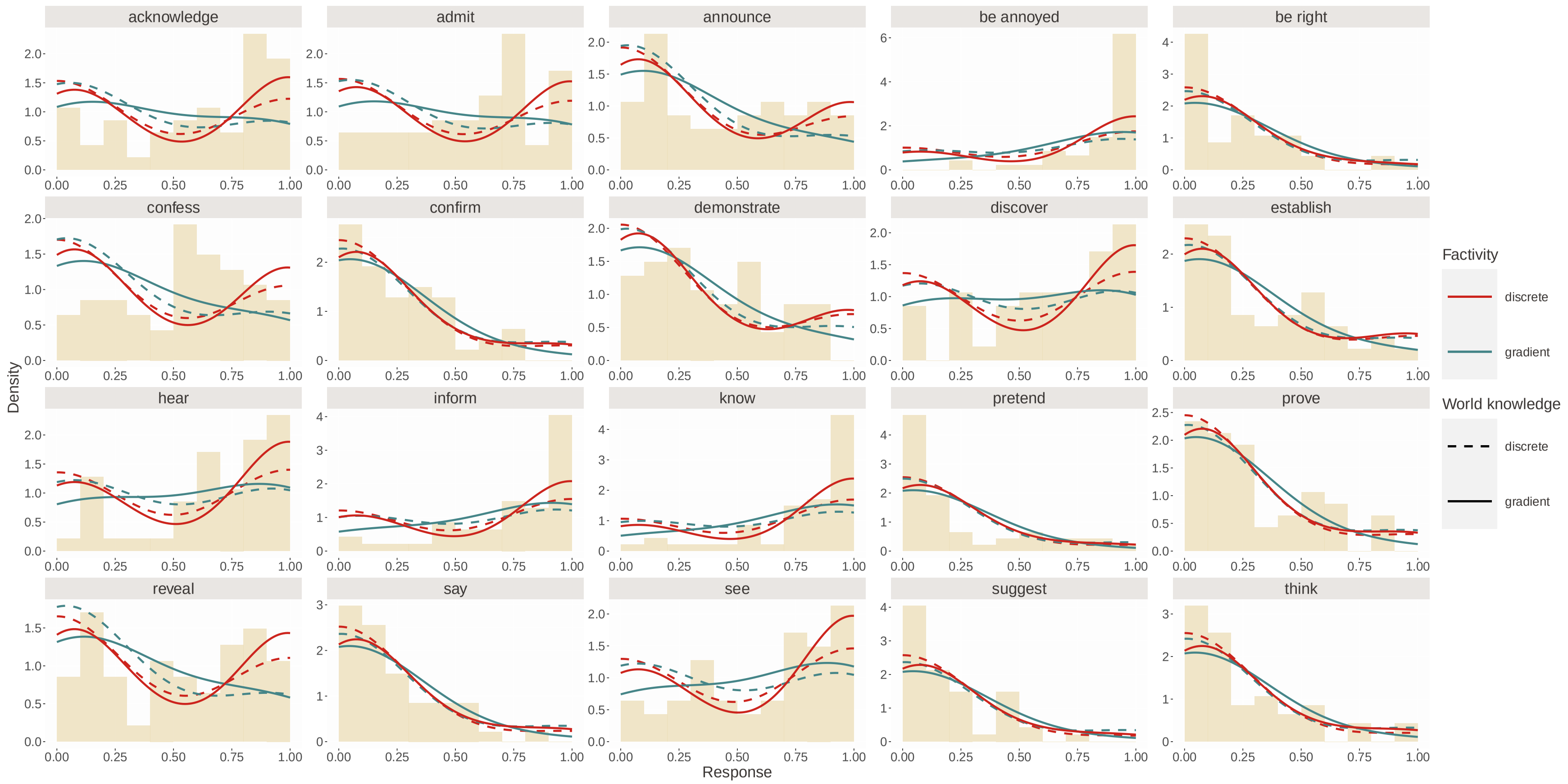


Figure 21: Posterior predictive distributions (with simulated participant intercepts) of all four model evaluations for all predicates in (3). Complement clause: *a particular thing happened*. Empirical distributions are represented by density histograms.

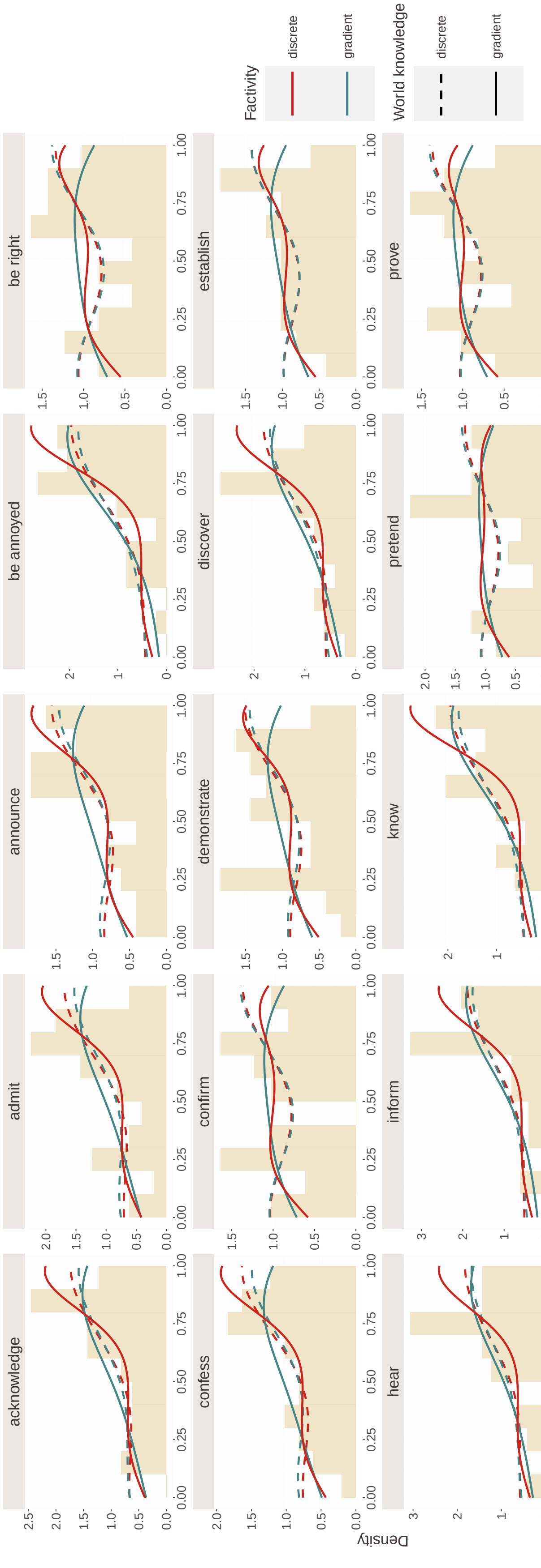


Figure 22: Posterior predictive distributions (with simulated participant intercepts) of all four model evaluations for all predicates in (3). Complement clause: *X happened.* Empirical distributions are represented by density histograms.