Pragmatic Reasoning and Semantic Convention: A Case Study on Gradable Adjectives[*]

Ming Xiang
University of Chicago

Christopher Kennedy
University of Chicago

Weijie Xu
University of California, Irvine

Timothy Leffel
University of Chicago

Abstract Gradable adjectives denote properties that are relativized to contextual thresholds of application: how long an object must be in order to count as *long* in a context of utterance depends on what the threshold is in that context. But thresholds are variable across contexts and adjectives, and are in general uncertain. This leads to two questions about the meanings of gradable adjectives in particular contexts of utterance: what truth conditions are they understood to introduce, and what information are they taken to communicate? In this paper, we consider two kinds of answers to these questions, one from semantic theory, and one from Bayesian pragmatics, and assess them relative to human judgments about truth and communicated information. Our findings indicate that Bayesian accounts can model human judgments about what is communicated better than they can model human judgments about truth conditions, but the performance improves if the Bayesian approach is supplemented with the threshold conventions postulated by semantic theory.

Keywords: Gradable adjectives, Bayesian pragmatics, context dependence, semantic uncertainty

[*] The data, codes and materials described in this manuscript can be found at: https://osf.io/nr6a4/. We are grateful to Thomas Sostarics for research assistance; to Judith Degan, Paul Egré, and two anonymous reviewers for their feedback on the ms.; and to the participants of the "Scales, degrees and implicature: Novel synergies between semantics and pragmatics" workshop for questions about, and reactions to, the ideas reported here. We are particularly grateful to thank Ciyang Qing for his feedback at various stages of this project and his generous support of our model implementations. This paper is based upon work supported by the National Science Foundation under Grant No. 1227144, and by a University of Chicago Division of Humanities Visiting Committee Faculty Research Grant.

1 Introduction

1.1 Gradable adjectives and threshold uncertainty

Gradable adjectives are predicative expressions whose semantic content is based on a scalar concept that supports orderings of the objects in their domains. For example, the gradable adjectives *long* and *short* order relative to length; *heavy* and *light* order relative to weight, and so on. Non-gradable adjectives like *digital* and *next*, on the other hand, are not associated with a scalar concept, at least not grammatically.

There are different formal characterizations of gradability in the literature, and of the difference between gradable and non-gradable adjectives, but one feature that all analyses agree on is that gradable adjectives are distinguished from their non-gradable counterparts in introducing (either lexically or compositionally) a parameter that determines a THRESHOLD of application, such that a predicate based on a gradable adjective holds of an object just in case it manifests the relevant property to a degree that is at least as great as the threshold. A predicate expression formed out of a gradable adjective therefore comes to denote a property only after a threshold has been fixed.[1] Comparatives, measure phrases, intensifiers and other kinds of degree constructions are examples of expressions that fix the threshold compositionally. For example, *two meters* in (1a) sets the threshold at two meters of length; *-er (= more) than this knife* in (1b) sets it to the length of the knife in question; *too ... to fit in the rack* in (1c) sets it to the maximum length consistent with fitting in the rack, and so forth.

(1) a. That pole is two meters long.
 b. That pole is longer than this knife.
 c. That pole is too long to fit in the rack.

Our concern in this paper is the interpretation of gradable adjectives in the morphologically unmarked POSITIVE FORM, which is illustrated by (2a-c).

(2) a. That pole is long.
 b. That knife is long.

[1] We use “threshold” here in a rather descriptive sense, as a cover term for that feature of a particular formal theory of gradable adjective meaning that is crucially involved in modeling variability in the extension of the predicate. The main point of divergence between formal theories of gradability has to do with whether the threshold is characterized as an actual argument of the adjective or adjectival projection, and if so, what its semantic type is, or whether it is a non-syntactic parameter of evaluation, subject to certain consistency constraints. (See e.g., Klein 1991, Kennedy 1999, Burnett 2016 and Qing 2020 for overviews of the different approaches and the syntactic and semantic issues at stake.) For the kinds of constructions we are interested in analyzing in this paper, which involve the meaning of the unmodified, “positive” form of the adjective, this distinction is irrelevant, as the subsequent discussion will make clear.

c. That rope is long.

The threshold of a positive form gradable adjective is not fixed compositionally by some other expression, and in the literature, it is typically said that, instead, the threshold is "determined by context." (See Lewis 2020a,b, Qing 2020 for good discussions of what exactly this amounts to.) And indeed, it is evident that the property expressed by a gradable adjective in the positive form is context dependent in a way that is consistent with the idea that the threshold can vary. (2a) might be judged true of a two meter long pole when it is lined up next to an array of smaller poles, but false of the very same pole when it is lined up next to an array of longer ones. Similarly, what we learn about the length of the pole from an assertion of (2a) is different from what we learn about the length of the knife from an assertion of (2b), or what we learn about the length of the rope from an assertion of (2c): a long pole is (normally) longer than a long knife, and is (normally) shorter than a long rope. This means that the contexts in which assertions of each of these different sentences are made determine distinct thresholds, such that we see variation in truth conditions — how long counts as *long* — and we draw different conclusions about the (minimum) lengths of the objects that *long* is predicated of.

There is an important difference between gradable adjective thresholds and the parameters relative to which the meanings of many other context dependent expressions are determined, however. In the case of, for example, the implicit internal argument of a noun like *resident* in (3a) or the implicit quantifier domain restriction in (3b), it is generally the case that successful instances of communication involve certainty about the semantic value of the relevant parameter.

(3) a. Are you a resident?
b. Everyone is here.

When a park ranger at the entrance of the Indiana Dunes State Park uses (3a) to determine whether to charge a visitor the regular fee or the lower fee for Indiana residents, it is clear that the semantic value of the implicit argument of the noun is the state of Indiana. Likewise, when the chair of the Linguistics Department says (3b) at the beginning of a meeting to vote on a colleague's tenure case, it is clear that the value of the quantificational domain restriction is the set of individuals designated to participate in the vote. A failure to understand these utterances in these ways results in a failure of communication in these contexts.

In contrast, in utterances involving positive form gradable adjectives, it is generally not the case that there is certainty about the value of the threshold. This is shown most clearly by the fact that gradable adjectives have borderline cases: objects about which we cannot say whether the predicate applies, even if we know the relevant facts about the objects themselves and the relevant facts of the conversational context. For example, if we go to a garden shop with the goal of purchasing a pole to support a small tree, and the salesperson presents us with an array of poles with clearly marked lengths ranging from 1 meter to 3 meters in 1 centimeter increments, there will be some poles about which we would be willing to assert (2a), some about which we would be willing to assert its negation, and some about which we would be willing to assert neither (2a) nor its negation. If there were certainty about where the threshold for length is in this context, this would not be the case: compare (2a) to the sentences in (1), each of which we would be willing to assert or deny about any of the poles, provided we also know the lengths of the knife and the rack.

Gradable adjectives with inherently context dependent and uncertain thresholds, such as *long, heavy* and *big*, are often referred to as RELATIVE gradable adjectives. But not all gradable adjectives have inherently uncertain thresholds. Alongside relative adjectives stands a class of ABSOLUTE gradable adjectives, which can manifest threshold uncertainty, but which also have uses in which there is relative certainty about the threshold (Unger 1975, Pinkal 1995, Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007, Toledo & Sassoon 2011, Lassiter & Goodman 2013, Qing & Franke 2014, Qing 2020). The adjectives *straight, empty* and *flat* in (4), for example, have uses in which they are true of their arguments just in case the objects in question have maximal degrees of the relevant property, and false otherwise.

(4) a. That pole is straight.
b. That theater is empty.
c. That countertop is flat.

Similarly, the adjectives *bent, open* and *striped* all have uses in which they are true of their arguments just as long as they have a non-zero degree of the relevant property, and false only if they lack the property entirely.

(5) a. That pole is bent.
b. That door is open.
c. That shirt is striped.

Note that the claim is not that there is no uncertainty about the thresholds for absolute adjectives at all; rather it is that that they have uses in which there is a high degree of certainty about the threshold, and that they show a correspondingly more limited range of context dependence than relative adjectives. For example, it is common to characterize a theater with a small but non-zero number of occupied seats as empty, though it would be strange to describe a half-full theater that way, and it is often fine to describe a pole with only a small amount of bend as straight or not bent, but not one with a ninety degree bend. Such "imprecise" uses of absolute adjectives introduce uncertainty about thresholds, and whether they are acceptable is a matter of context. A disgruntled theater owner could appropriately describe a theater with just a few occupied seats as empty when talking to the manager of a band that failed to draw an anticipated crowd, but it would be inappropriate for the theater owner to describe the same theater as empty when speaking to a detective who was interested in finding out whether a murder suspect might have been in the audience. Similarly, it would be natural for the owner of a dive bar to describe their pool cues as straight or not bent even if they are slightly bent. But it would be inappropriate for an engineer to describe an axle they are creating for a sensitive piece of machinery as straight or not bent when it has the same degree of bend. It is in these latter, “precise” contexts, that we see certainty about the threshold: it corresponds to a maximum or minimum value on the relevant scale.

1.2 Two theories of thresholds

Threshold uncertainty leads to two questions about the semantics and pragmatics of gradable adjectives in particular contexts of utterance:

(S) What are the truth conditions of such utterances?

(P) What is the information communicated by such utterances?

These are questions that one can of course ask about all sorts of expressions, and the answers can have non-trivial theoretical significance. For example, Grice (1975) and subsequent work in the Gricean and neo-Gricean tradition provide different answers to (S) and (P) in their accounts of upper-bounded interpretations of weak scalar terms, while approaches that derive such interpretations from the compositional contribution of an exhaustification operator in the syntax (e.g., Fox 2007, Chierchia et al. 2012, etc.) are committed to giving the same (or nearly the same) answers. The case of gradable adjectives is particularly interesting, because the fact of threshold uncertainty suggests that, except perhaps for the special case of absolute adjectives on precise uses, it is impossible to provide an exact answer to (S). And yet, the fact that such expressions are systematically and successfully used to communicate information about the degrees to which objects manifest scalar properties shows that this does not present a problem for answering (P). In the following sections, we discuss two theories of threshold determination for gradable adjectives, which in effect constitute answers to (S) and (P), respectively.

1.2.1 Semantic accounts

The relative/absolute distinction is based on the interpretation of the positive form: whether the threshold is inherently uncertain, or whether it tends to correspond to a maximum or minimum value. But whether an absolute interpretation is even an option depends on a lexical semantic feature that varies across gradable adjectives: whether they encode scalar concepts that are based on open or closed scales, i.e. scales which respectively lack or include minimal or maximal degrees. This can be diagnosed by looking at acceptability with certain types of modifiers (Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007, Syrett 2007, Solt 2012). The modifier *completely*, for example, introduces the entailment that an object has a maximal degree of a gradable property, and so combines only with adjectives that use scales with maximum values, while the adjective *slightly* entails that an object exceeds a minimum degree, and so selects for adjectives that use scales with minimum values. As the following examples show, there is a correlation between the relative/absolute distinction and scale structure: absolute adjectives have closed scales; relative adjectives have open scales.[2]

(6) a. completely straight/empty/flat
b. # completely long/heavy/big

(7) a. slightly bent/open/striped
b. # slightly long/heavy/big

This correlation between scale structure and the relative/absolute distinction has given rise to a family of accounts that link threshold determination to the lexical semantics of the predicate (Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007, Toledo & Sassoon 2011, Burnett 2016, Qing 2020). There are differences in implementation between these accounts, but they share the general feature that closed-scale adjectives can be conventionally associated with endpoint-oriented thresholds, giving rise to absolute truth conditions. This is not an option for open scale adjectives, in contrast, since they use scales that lack maximal or minimal values, and so the value of the threshold — and the truth conditions — must be "fixed by context."

1.2.2 Bayesian pragmatic accounts

Lassiter & Goodman (2017, 2013) (LG) develop a Bayesian model of communication with gradable adjectives that starts from what is arguably the null hypothesis about their semantics: since both relative and absolute adjectives combine with

[2] The examples in (7b) are crucially unacceptable on interpretations that are parallel to the most prominent interpretations of the examples in (7a), which would be paraphrased as "a slight amount of length/weight/size." These examples can have a different kind of interpretation, paraphrasable as "slightly *too* long/heavy/big," i.e. as expressions of slight excess. But in such cases the semantics of excess provides a minimum standard for the modifier to interact with, namely the minimum degree that counts as excessive for the relevant purpose. Similarly for comparatives (*slightly longer than*) and even for positive form adjectives with non-vague standards (Solt 2012).

expressions that compositionally manipulate thresholds, and since both can have context dependent interpretations in the positive form, there is no special lexical semantic feature (such as differences in scale structure) that differentially determines how thresholds are fixed in context. Instead, thresholds are always uncertain, and any truth conditional or communicative differences between the two classes of adjectives is to be explained in terms considerations outside of the semantics proper.

These considerations, according to Lassiter and Goodman, involve a general pragmatic strategy for determining what is communicated in the presence of semantic uncertainty. The LG model is implemented within the Bayesian Rational Speech Act (RSA) framework (Goodman & Frank 2016), which models language communication as a recursive process of pragmatic reasoning between rational agents. A simple version of this imposes some bound on the recursive reasoning process. A pragmatic listener L_1, upon hearing an utterance, updates their probabilistic understanding of the world states by reasoning about what a pragmatic speaker S_1 could have chosen as their utterances. The pragmatic speaker S_1 makes a choice on the utterance by reasoning about a literal listener L_0, who only considers the compositional semantics of the utterance without any pragmatic reasoning. Following the general RSA approach, the LG model captures the interpretational differences between different classes of adjectives as a matter of pragmatic inference.

Under the LG model, a pragmatic listener, upon hearing an utterance u containing a gradable adjective — E.g., (2a), *“That pole is long”* — simultaneously infers both the length of the object ℓ and the relevant threshold θ. The formal definition of this inference follows Bayes’ Rule:

(8) $P_{L1}(\ell,\theta \mid u) \propto P_{S1}(u \mid \ell,\theta) \times P(\ell) \times P(\theta)$

The posterior joint probability of ℓ and θ, for a pragmatic listener, given an utterance u, $P_{L1}(\ell,\theta \mid u)$, is determined by three factors: the prior probability for ℓ, the prior for θ, and the probability that a speaker would choose to utter u given ℓ and θ. The prior for ℓ comes from the listener’s prior beliefs about the length distribution of particular categories in the world, for example, the length distribution of garden poles at Home Depot. This is mostly determined by a listener’s world knowledge. The prior for θ is assumed to be uniform. That is to say, a listener does not need to hold any background assumption about any particular threshold. They update their beliefs about the threshold in a particular context upon hearing the utterance. The probability of a speaker choosing to utter the adjective to describe the object can be computed using the equation in (9):

(9) $P_{S1}(u \mid \ell,\theta) \propto exp(\lambda(informativity(u,\ell,\theta) - cost(u)))$

A speaker, in the simplest scenario, could choose to stay silent or make an utterance. The probability of them making an utterance — saying *“that pole is long”* instead of saying nothing — is determined by the utility of the utterance, which in turn reflects a trade-off between its *informativity* for the listener and the *cost* of producing it for the speaker. The informativity of an utterance is defined over the posterior probabilities a *literal listener* holds about ℓ in situations in which u is true:

(10) a. $\mathit{informativity} = \mathit{log}(P_{L0}(\ell \mid u, \theta))$

b. $P_{L0}(\ell \mid u, \theta) = P(\ell \mid [\![u]\!]^{\theta} = 1)$

A speaker therefore evaluates the informativeness of their utterance by conditioning on its truth conditions; in the case of a positive form adjective like *long*, the truth conditions require that $\ell \geq \theta$. The cost of an utterance is an intuitive notion, but there is actually no fully predictive theory of what its exact value should be, and we follow the common practice of treating it as a free parameter that is inferred from the empirical data. A second free parameter in (9) is $\lambda > 0$, which quantifies the degree of rationality of the speaker model, that is the degree to which utility is maximized.

It is important to point out that, after putting equations (8) to (10) all together, the (listener's) threshold value is completely determined pragmatically, at equation (8). Although the truth conditions of an utterance containing a positive form adjective make reference to thresholds, listeners have no a priori commitment about them (assuming a uniform prior distribution for θ). They can infer the values of thresholds by considering, for all possible threshold values and all possible messages intended by the speaker — in this example, all possible values for ℓ — how likely it is that a speaker would have uttered the adjective. It is only at the end of this iterative reasoning process that a listener derives an updated posterior belief about the distribution of θ, as well as a posterior distribution of ℓ. It may seem a bit counter-intuitive that prior to the interpretation of an utterance, a listener has no commitment about thresholds for any kind of adjectives. But this was considered by Lassiter and Goodman as a desirable feature of the model since it supports a fully general account of the difference between relative and absolute adjectives that is based on differences in prior beliefs about how objects distribute relative to various scalar concepts.

For example, assume as above that (2) is used to describe a garden pole at Home Depot; here the relevant prior for lengths is based on the listener's beliefs about the lengths of similar poles — the comparison class — which we may assume to have an approximately normal distribution. The pragmatic reasoning process is crucially sensitive to the informativity of the literal semantic meaning of the utterance — that the pole in question has a length greater than θ — for various values of θ. The lower the threshold is, the more likely it is that an arbitrary pole has at least that much length; and the higher the threshold is, the less likely it is that an arbitrary pole has that length. As a result, a low value for θ (e.g., one that makes the utterance true of 75% of the poles in the comparison class) will be assigned low probability, because the resulting meaning would be too weak, while a high value for θ (e.g., one that makes the utterance true of only 1% of the poles) will also be assigned a low probability, because the resulting meaning would be too strong. In theory, the output of the LG model in a simple case like this is a posterior probability distribution for thresholds that is shifted upwards from the prior degree distribution over the relevant comparison class, and a posterior probability distribution for the length of the target of predication that is shifted still further up the scale, and (2) is (correctly) predicted to communicate something roughly equivalent to "the length of that pole is significantly greater than the average length of poles in the comparison class."

In the case of an utterance involving an absolute adjective like (4a) *"That pole is straight,"* the pragmatic reasoning process works in exactly the same way, but delivers a different kind of output because the prior distribution for degrees of pole-straightness is different from the prior distribution for degrees of pole-length. While the latter is (plausibly) normal, the former is not; instead there is substantial probability mass at the upper end of the ordering: we tend to encounter a lot of straight poles. The output of the model described in Lassiter & Goodman 2013 in such a case is a high posterior probability that the threshold for *straight* is selected from a narrow range of values near the scalar maximum, and a correspondingly high degree of posterior probability that that the straightness of the pole is at or near the maximum. (A minimum standard interpretation of *bent pole* can be derived from the same priors, given the assumption that antonym pairs lexicalize inverse ordering relations.) This is why absolute adjectives give rise to the appearance of fixed thresholds, compared to relative adjectives: in both cases, there is uncertainty about the threshold and corresponding uncertainty about the degree to which the target of predication possesses the relevant property, but in the case of absolute adjectives, this uncertainty is significantly reduced.

A second, more speaker-oriented Bayesian model of gradable adjective interpretation is proposed in Qing & Franke 2014 (QF), which shares some features with the LG model, but critically diverges in its conceptualization and technical implementation of the notion of threshold. Instead of making the threshold purely the outcome of a pragmatic reasoning process that guides a particular linguistic exchange situation, thresholds for adjectives are viewed as linguistic conventions learned in a community to achieve optimal communicative efficiency between a speaker and a listener. In this sense, the QF model aims to capture how a community solves an optimization problem to form good semantic systems, whereas the LG model focuses on capturing the pragmatic reasoning that takes place "on the fly" between interlocutors. In the QF model, when a speaker makes a choice between uttering an adjective or saying nothing, they already have probabilistic knowledge $\Pr(\theta)$ about the distribution of θ for a given adjective. A listener then updates their belief about the relevant property of the target object, conditioned on the speaker's utterance, via Bayes' rule in the following way (equation (9) in Qing & Franke 2014):

(11) $P_L(\ell \mid u) \propto P_S(u \mid \ell, Pr(\theta)) \times P_L(\ell)$

Comparing the QF listener in (11) with the LG listener in (8), the QF listener does not need to infer the values for θ on the fly. This is made possible because the speaker model $P_S(u \mid \ell, Pr(\theta))$ is assumed to already have the knowledge of $Pr(\theta)$. For the LG listener in (8), on the other hand, θ is a free variable that is passed up from the literal listener in (10b), and the value of this variable is only resolved when the pragmatic listener jointly infer both θ and the length of the target object in (8).

The critical task for the QF model, then, is to explain and derive the speaker's probabilistic knowledge about the threshold distribution $Pr(\theta)$. We refer interested readers to Qing & Franke (2014) for a detailed discussion on how $Pr(\theta)$ is derived, but we want to highlight two important features. First, in an evolutionary perspective, threshold distribution is a semantic convention derived under the communicative pressure that the linguistic community wants to settle on thresholds which, on average, will help listeners most successfully pick out the correct degree that a speaker intends to convey when choosing to use a positive form gradable adjective. The best θ is the one that, after a listener updates their prior belief based on the utterance of a speaker, they would have the best chance to arrive at the intended degree. The communicative success of a threshold in the long run is measured in Qing & Franke (2014) by the *expected success rate* of θ. The expected success of a threshold, combined with considerations about production effort (i.e. a cost parameter), determine the utility function of θ. The probability distribution of θ $Pr(\theta)$ is finally computed based on the utility function.[3] Without going through the technical details, the upshot is that different types of adjectives may come to be conventionally associated with different kinds of thresholds — minimal, relative, maximal — in virtue of the fact that these turn out to be the optimal thresholds that achieve the best communicative efficiency.

This points to the second crucial feature of the QF model, which is that although $Pr(\theta)$ emerges as a semantic convention from the evolutionary component of the model, such that relative and absolute adjectives may be conventionally associated

[3] In Qing & Franke 2014, the expected success rate of θ, the utility function of θ, and $Pr(\theta)$ are defined in the following way (see equations (6-8) in the paper):

(i) $ES(\theta) = \int_{-\infty}{}^{\theta} P(\ell)P(\ell \mid u_0, \theta)d\ell + \int_{\theta}{}^{\infty} P(\ell)P(\ell \mid u_1, \theta)d\ell$

(ii) $U(\theta) = ES(\theta) - \int_{\theta}{}^{\infty} P(\ell) \cdot cd\ell$, where c is a cost parameter

(iii) $Pr(\theta) \propto exp(\lambda \cdot U(\theta))$, where λ is a parameter of rationality with different kinds of thresholds, it does so ultimately in virtue of differences in prior beliefs about distributions of the objects in the comparison class along the relevant scalar dimensions, for example beliefs about the lengths or straightnesses of the sorts of poles found at Home Depot. The distinction between relative and absolute adjectives emerges because different degree scales, e.g., open vs. closed scales, constrain the priors, which in turn influence how people form their knowledge about the optimal thresholds. Other information, such as world knowledge, can also constrain priors as well. In this sense, QF shares with LG the feature that thresholds are grounded in prior beliefs about how objects in the world distribute along various dimensions, instead of in lexical semantic features of gradable adjectives.

1.3 The current study

As noted above, the semantic and Bayesian pragmatic theories of thresholds described in the previous sections in effect constitute theories of how to answer the following two questions, respectively:

(S) What are the truth conditions of such utterances?

(P) What is the information communicated by such utterances?

Semantic theories answer (S) by providing conventions for fixing the value of the threshold; Bayesian theories answer (P) by providing posterior degree probabilities. The different approaches therefore make different kinds of predictions about behavior relating to (S) and (P).

Semantic theories, which are geared to answer (S), predict that truth value judgments for relative adjectives should vary with context, and should not be categorical, due to uncertainty. Truth value judgments for absolute adjectives, in contrast, should largely be categorical and context invariant. Such theories say very little about the answer to (P), however, and what they do say appears to be wrong. In the case of relative adjectives, in the absence of any theory of how thresholds are "fixed by context," it is difficult to say exactly what a semantic theory predicts about what is communicated. In the case of absolute adjectives, the theory provides clear answers to (P), but the wrong ones: the use of a maximum threshold adjective like *straight* should communicate that an object has maximal straightness, but such utterances generally communicate something weaker; the use of a minimum threshold adjective like *bent* should communicate merely that an object has some amount of bend, but such utterances generally communicate something stronger.

The Bayesian pragmatic accounts, in contrast, do not suffer from these problems because they are designed to answer (P), not (S). The pragmatic listener, upon hearing *"that pole is long/straight/bent"* aims to update his/her probabilistic belief about the pole's degree of height, straightness, or bend. On the other hand, these approaches only make direct empirical predictions for behaviors about posterior degree judgments: they do not directly speak to truth value judgments, though they can be made to do so if they are supplemented with additional linking hypotheses (see more discussion about this in section 3.2).

A second difference between the semantic approaches and the Bayesian pragmatic approaches is that the latter but not the former critically makes use of language users' prior knowledge of degrees along various dimensions. Both the LG and the QF models aim to update a prior distribution of degrees to a posterior one, conditioned on a certain utterance. The prior degree distribution, in both models, plays an important role in shaping the posterior. One can make reasonable assumptions about the possible theoretical distributions for the prior. However, if the Bayesian approaches aim at providing a cognitively plausible mechanism for capturing human linguistic behavior, it is a non-trivial empirical question as to what kind of priors language users actually have access to.

In light of these considerations, the goals of the current study are twofold. Our primary goal is to elicit truth value judgments and posterior degree judgments from human subjects, and use these data to evaluate the predictions of the different approaches. We cannot ask whether semantic approaches make correct predictions about posterior degree judgments, since they are not designed to do so; but we can ask whether the Bayesian pragmatic approaches, when supplemented with a plausible linking hypothesis, make correct predictions about truth value judgments. A secondary goal, given the reliance of the Bayesian approaches on prior degree distributions, is to also elicit empirical priors from human subjects and use these to compute the model predictions, rather than relying on artificial priors, as in previous studies.

The remainder of the paper is organized as follows. Section 2 presents the results of three experimental tasks, which collect empirical (human) prior degree estimations, truth value judgments, and posterior degree estimations, respectively. Next, in Section 3, we use the empirical priors to generate predictions about truth values and posterior degrees for the LG and the QF models, and compare these to the human data that we collected. As we will show, using the empirical priors, the model predictions perform better at matching the human results for posterior degree estimations than for truth value judgments; the human truth value judgments correspond closely to the predictions of semantic theories. In an exploratory analysis in section 4, we show that by providing a Bayesian model with θ values that are consistent with the thresholds postulated by the semantic theories, the model still can maintain its performance on posterior degree estimations, suggesting that supplementing the Bayesian model with threshold conventions that are compatible with traditional semantic theories has the most successful empirical coverage. We also discuss the implications of these findings.

2 Empirical estimates from human participants

2.1 Experiment 1: Degree priors

2.1.1 Methods

Methodologically speaking, it is not obvious what would be the best experimental paradigm to elicit degree priors. Since our goal was to establish a probability distribution for the degrees that objects in a comparison class are believed to have relative to different scalar dimensions (e.g., the probability that an arbitrary garden pole has a length ℓ, for a range of lengths) *independent* of any facts about language users' experience with the words that are used to talk about these scales (e.g., *length*, *long*, *short*, etc.), we decided that no such words should be used to elicit degree priors. Instead, we used an aggregated judgment of likelihood as a proxy for prior degree probability: we presented subjects with a set of items that were identical as much as possible in all respects except for the degree to which they manifested a particular scalar property (degree of length, degree of height, degree of bend, degree of fullness, etc.), and asked subjects to choose the single object from the group that they believed to be "most likely."

Partly as a reality check on our methodology, we divided our stimuli into two categories of objects, which we expected to give rise to different patterns of degree priors. The first category, which we call *artifacts*, consisted of objects that are common in daily life, such as candles, pillows, nails, etc. Since people have relatively rich and varied experience with these kinds of objects in different kinds of contexts, we expect that they will be more likely to have fairly fine-grained prior beliefs about how these objects distribute along our scalar dimensions of interest. The second category of objects, which we call *shapes*, consisted of abstract, geometric shapes that people are likely to have only occasional experience with, and in more limited contexts, such as triangles or cylinders of the sort that appear in mathematics textbooks. We expect subjects to have less fine-grained priors for such objects, or at least more categorical ones, for the scalar dimensions under examination. If the judgments we collect for artifacts and shapes turn out to be distinct in this way, we will have reason to believe that our methodology is reasonably capturing priors; if the judgments are not distinct, we will have reason to think that it is not.

A second reason for introducing the shapes/artifacts distinction is that it allows us to test the predictions of the Bayesian models in a fine-grained way. If there are prior differences between the two types of objects, the Bayesian models would predict that the same adjective would have different interpretations depending on whether it is applied to a shape or an artifact object. The traditional semantic approach would not make this prediction. In fact results from a study by Foppolo & Panzeri (2011) indeed support the former prediction, who showed that experimental subjects were more likely to judge a sentence of the form *"x is adj"*, where *adj* is a maximum absolute adjective, as true of an object with a high but non-maximal degree of the property denoted by *adj* when the object was an artifact than when it was a shape, demonstrating the impact of world knowledge prior on the interpretation of absolute adjectives. Observations like this lend support to the Bayesian approach and raise questions to the categorical treatment of absolute adjectives in the traditional semantic approach. The current study therefore uses the artifact-shape distinction to first reproduce Foppolo and Panzeri's results in a truth value task. We then extend the investigation to the estimation of posterior degrees, and we also ask whether the Bayesian models can capture the differences (if any) between the two categories that are observed in the human responses.

Participants Ninety-seven participants completed the study on IbexFarm; all were self-reported native English speakers recruited from MechanicalTurk (mean age: 34; 40 females). All participants were located in the US. Each participant was compensated at a rate of $10/hour. The experiment was approved by the University of Chicago IRB board. All the experiments reported in this paper had the same participant recruitment procedure.

Materials Forty-eight sets of images were created, 24 for artifacts and 24 for shapes. Each image set consisted of five items that differed in the degrees to which they manifested a particular scalar dimension. The dimensions were selected so that they could later be associated with members of pairs of antonymous adjectives in the truth value and posterior degree experiments, as described below. Examples of image sets are shown in Figure 1.

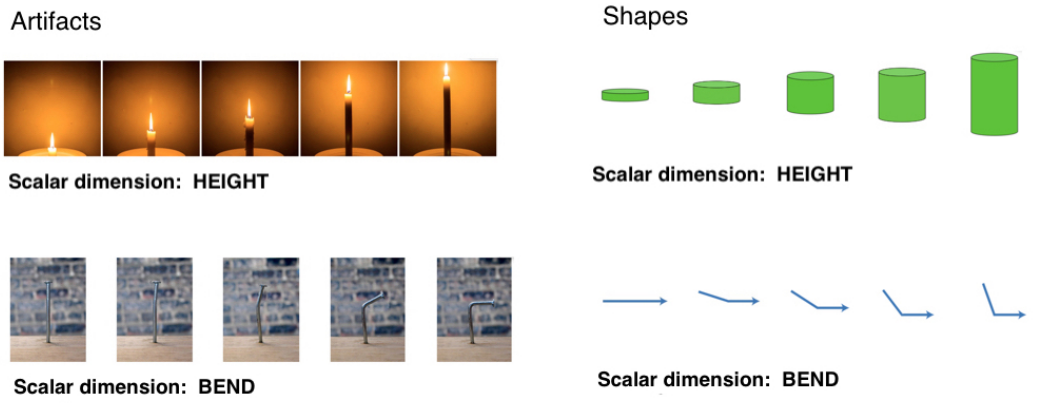


Figure 1: Sample image sets

Procedure The artifacts (24 sets) and shapes (24 sets) were tested in the same experiment. For each image set, participants were presented with the images from the 5 scale positions, and were asked *"Which of these is the most likely?"* No specific adjectives were mentioned for any stimuli. For each trial, participants were allowed to choose only one of the five objects in the image set. Figure 2 shows an example of the trials that participants saw. The priors elicited in this way on each image set were later associated with a pair of antonyms in Experiment 2 and 3. For instance, in Figure 2, the priors elicited for the arrow image set are later associated with judgments about *bent* and *straight* arrows, and the priors elicited for the candle image set are later associated with judgments about *tall* and *short* candles.

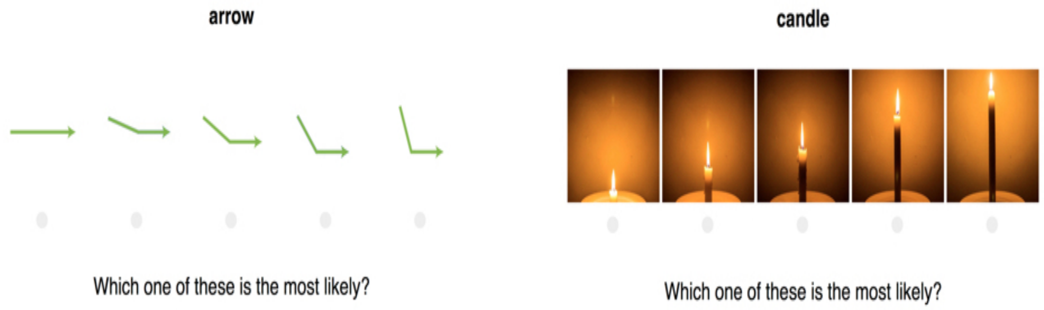

arrow
Which one of these is the most likely?
candle
Which one of these is the most likely?

Figure 2: Sample stimuli for Experiment 1

2.1.2 Results and discussion

For each image-set, we first averaged responses from participants to get the relative frequency that each scale position was chosen as the *most likely*. We used the relative frequency measure (i.e. a proportion) as a proxy for the probability distribution of priors. The result from this by-item calculation, i.e., the prior distribution over the five scale points on each image set, is the input to the Bayesian models when we discuss the modeling results in Section 3. In the prior-elicitation experiment, we intentionally did not use any adjective. But in Figure 3, for visualization purpose only, we show a "by-adjective" result for the proportion of selection of each scale point. This was done through averaging over the items that shared the same adjective when we later elicited truth judgments and posterior degree estimations in Experiments 2 and 3. For example, since the *candle* image set and the *book stack* image set both appeared with the adjective *tall* in the later experiments, their results from Experiment 1 were averaged together. For each adjective, scale position 1 corresponds to the object that has that property to the least degree, and scale position 5 corresponds to the object that has that property to the greatest degree. For example, scale position 1 for *big* is the least big (smallest) object, and scale position 1 for *small* is the least small (biggest) object. In most cases, the priors shown in Figure 3 for antonym adjectives are mirror images of each other, since judgments about an antonym pair in later experiments (e.g., judgments involving *big* and *small*) were made about the same images. However, a small number of adjectives were used in different sets of antonym pairs — for example, *short* was paired with *long* and with *tall*, and *straight* was paired with *bent* and with *curved* — with different image sets in each case. As a result, the prior distributions shown in Figure 3 for these antonym pairs are not precise mirror images of each other. For more details about the adjectives used and the distribution of them, see the materials sections for Experiment 2 and 3.

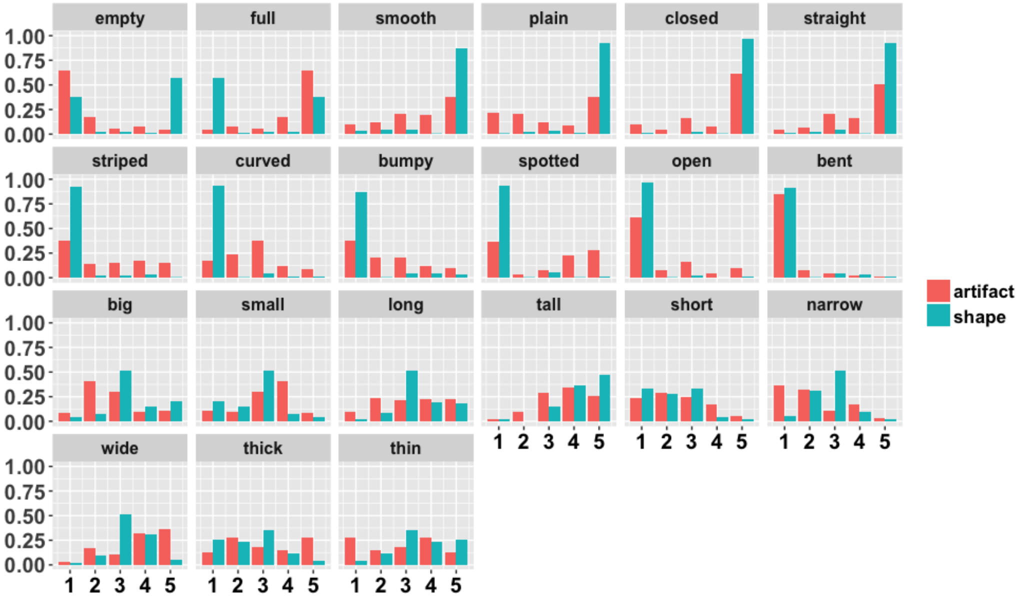


Figure 3: Experiment 1: Likelihood judgments as proxies for degree priors. Bars represent proportion of selection for each scale position for the image sets used for each adjective in Experiments 2 and 3. **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

Two features of the results in Figure 3 lead us to believe that our method for eliciting priors, which takes aggregated judgments about likelihood as standing proxy for prior beliefs about degree distributions, is reliable. First, the likelihood proportions differ by adjective class in the same way that the Bayesian pragmatic accounts of the absolute/relative distinction hypothesize the degree prior distributions to differ: absolute adjectives tend to show greater probability mass at scalar endpoints, while relative adjectives have more spread-out distributions. And second, the pattern of responses for shapes and artifacts are different in the way we anticipated: artifacts tend to have a less categorical distribution than shapes, in particular for the dimensions corresponding to absolute adjectives.

That said, since there is relatively little prior work geared towards collecting empirical priors for gradable adjectives, it is worth considering some potential problems with our task and some alternative methodologies before moving on.[4] One potential concern is that the prompt *"Which of these is the most likely?"* was unnatural, and that it would have been better to ask a more specific question such as *"Which of these is most likely to occur in the world?"* We acknowledge that the prompt we used was unnatural, but given the range of image types that we presented to subjects — from geometric cylinders of different sizes to photographs of actual garage doors with different degrees of openness — we felt it was better to use the more generic prompt rather than other alternatives that were more specific.

A more significant concern is that in aggregating judgments about which object is "most likely" across participants, our method might not accurately generate something that corresponds to an individual subject's prior probability distribution. In particular, in asking subjects to choose the single "most likely" object, our method might have biased participants towards more salient choices and inappropriately amplified differences between the the highest prior probability and the others. First we note that our method is conceptually similar to the *give-a-number* task discussed in Franke et al. (2016), which compared three different tasks to elicit priors. In a *give-a-number* task, participants answered questions like "how many minutes do you think she spent commuting yesterday" by using a slider to choose a number, after reading a context sentence about a person commuting to work. Our method likewise asked people to choose a value from a range of values; the major difference is that whereas the study in Franke et al. (2016) could explicitly ask a question about numerosity, we could not explicitly ask for a degree because we did not want to use the scalar term in the prompt. In addition, our participants had to choose among just five discrete points whereas in the previous study participants could choose from a much larger set of options, since it was easier to implement a larger set of options on a numerosity scale. Nonetheless, Franke et al. (2016) showed that the measures obtained from the *give-a-number* task are at least consistent with the measures from the other two tasks they conducted.

[4] We are grateful to the associate editor and anonymous reviewers for raising these questions.

Another promising task discussed in Franke et al. (2016) is the *binned histogram/mean slider rating* task. This task has been successfully used in a number of other studies that examine the interpretation of number terms (Kao et al. 2014, Schöller & Franke 2017). We therefore conducted an additional experiment using a version of this task. In this task, instead of asking participants to only select one single object, we asked participants to use a slider to rate every object on the scale according to how likely it is in the world. The mean slider ratings for each object were then normalized to derive a probability distribution over the five objects in the same object set. The study was carried out on IbexFarm using a slider controller developed in Chen & van Tiel (2021). More details of this experiment are provided in the Supplementary Material (Section S2). Importantly, the results we obtained with this method resulted in priors that were largely uniform across all scale types and so showed no sensitivity neither to the shapes/artifacts distinction nor to the distinction between different adjective types.

We will leave it as an open question why the mean slider rating task is not sensitive to the current experimental manipulations. In general, more future work is needed to establish the validity of different prior elicitation methods across a diversity of contexts. As discussed in Franke et al. (2016), ideally we would want a prior elicitation method that can provide reliable information for different empirical domains and at the same time is easy to understand by the participants, as well as easy to implement and analyze.[5] For the current purpose, since the method in Experiment 1 obtained reasonable priors that showed sensitivity to our experimental manipulations, we remain confident that our task is reliable for collecting empirical priors, and we will use the prior information estimated in Experiment 1 to test the predictions of the Bayesian models in Section 3.

2.2 Experiment 2: Truth value judgments

Our second experiment elicited human truth value judgments about sentences in which gradable adjectives were predicated of the same objects that were used to collect empirical priors in Experiment 1.

[5] One of the anonymous reviewers suggested a different task, which would ask participants to distribute 100 points to the five objects, proportional to how likely they think each one was (see a similar task in Frank & Goodman (2012) with a smaller number of objects). This task is conceptually appealing, but it is potentially a demanding task for the participants, since on every trial they have to perform an arithmetic calculation to add up 5 numbers to be exactly 100. The task demand may lead to more errors and could also encourage task-specific strategies.

2.2.1 Methods

Participants Experiment 2 was conducted on IbexFarm, with all participants recruited from MechanicalTurk. A total of 116 self-reported native English speakers participated, with 58 in the artifact group (mean age: 34; 27 females) and 58 in the shape group (mean age: 33; 21 females).

Materials Experiment 2 used the same 48 sets of images that were used for the elicitation of prior degree estimations in Experiment 1, with 24 artifact image sets and 24 shape image sets. Each image set was then paired with two adjectives that are antonyms with each other. For example, the example stimuli in Figure 4 could be presented together with either the adjective *striped* or the adjective *plain*. This resulted a total of 96 items, with 48 artifact and 48 shape items. Artifact and shape items were tested separately for two different groups of participants.

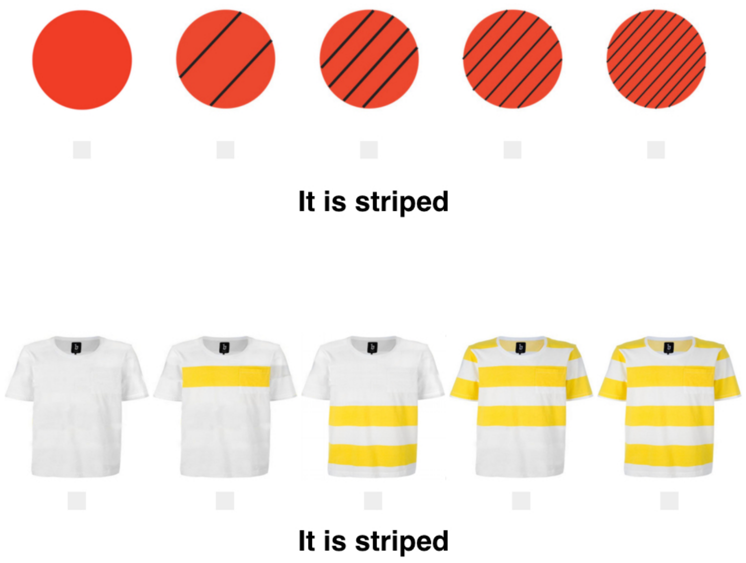


Figure 4: Sample stimuli for Experiment 2: Truth value judgments

For the 48 artifact items, a total of 21 adjectives were used in the study, among which were 6 maximum adjectives (*straight*, *closed*, *plain*, *smooth*, *empty*, *full*), 6 minimum adjectives (*curved*, *open*, *striped*, *spotted*, *bent*, *bumpy*), and 9 relative adjectives (*big*, *small*, *long*, *tall*, *short*, *narrow*, *wide*, *thick*, *thin*). These adjectives were organized into 12 pairs of antonyms: full-empty, bent-straight, big-small, wide-narrow, striped-plain, tall-short, open-closed, curved-straight, spotted-plain, thick-thin, long-short, bumpy-smooth. Each pair was used twice, i.e. each pair was used for two different image sets, resulting in 2 (adjectives in a pair) x 12 (antonym pair) x 2 (image sets with the same antonym pair) = 48 items. The adjective material for the 48 shape items was constructed in exactly the same way as the artifact items. The full set of stimuli items could be found in Supplementary Material (S1).

Procedure The artifact and the shape items were tested on two separate groups of participants, but the procedure for each group was identical. For each trial, participants were told that they would see a series of images and a sentence, and that they should click on the checkbox beneath the image or images that they believed the sentence "appropriately describes," a judgment that we take to stand proxy for truth value judgments. Participants could choose multiple images in one trial. Example stimuli are shown in Figure 4. Since each image set was paired (separately) with each member of an antonym pair, the 48 shape/artifact items were distributed in a latin-square fashion such that the same participant did not see both adjectives that were paired to the same image set. For example, one participant would only see the sentence *It is striped* for the example stimuli presented in Figure 4, and another participant would only see *It is plain* for the same images. Each participant, therefore, only saw 24 trials total.

2.2.2 Results and discussion

Figure 5 shows, for each adjective class (absolute maximum, absolute minimum, relative), the proportions of accepting a given utterance as true of an object at a given scale position. In Figure 6 we also present results for each individual adjective tested. For statistical analysis, we used the R package "brms" to fit a Bayesian mixed effects logistic regression model predicting true over false judgments (Bürkner 2017). All the models used weakly informative priors for the prior distribution of parameters. The first model included the effects of image type (artifacts vs. shape), adjective class (relative, maximum, and minimum) and scale positions, as well as their interactions, as the fixed effects, and the model also included by-participant, by-image_set, and by-adjective random intercepts, as well as the by-participant random slopes for adjective class and scale position, by-image_set random slope of scale position, and by-adjective random slopes for image type and scale position. The predictor image type was sum coded at (artifact 1, shape -1). Adjective class was treatment coded with the relative adjectives as the baseline level, such that the maximum and minimum adjectives are compared to the relative adjectives in the model output. Scale position was coded as a continuous predictor and was centered before entering into the model[6]. The estimations obtained from a Bayesian hierarchical model include the mean, the standard error (SE), and the lower and upper bounds of the 95% credible interval (CrI) of the posterior distribution for each parameter of interest. The 95% CrI can be interpreted as there is a 0.95 probability, given our data and prior assumptions, that the true population mean of the relevant parameter lies within this interval. We use this interval as our primary metric for drawing statistical inferences. In particular, if the interval excludes 0, it can be considered as providing support for an effect.

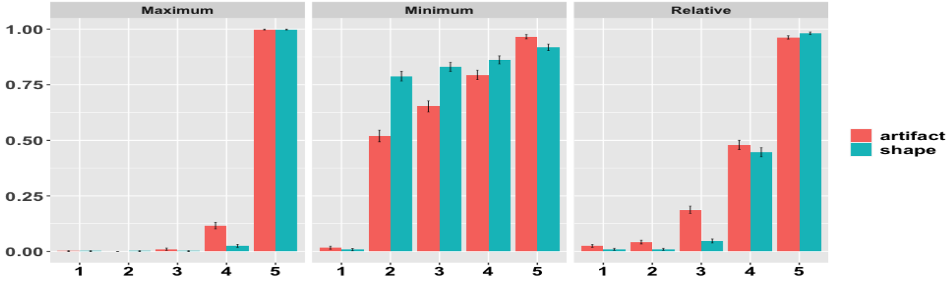


Figure 5: Experiment 2 Truth Value Judgment: Percentage of positive responses at each scale position for each adjective class. Error bars indicate standard errors.

[6] This is likely an idealization since, for example, one-unit increase from position 1 to 2 is probably not totally comparable to that from 2 to 3.

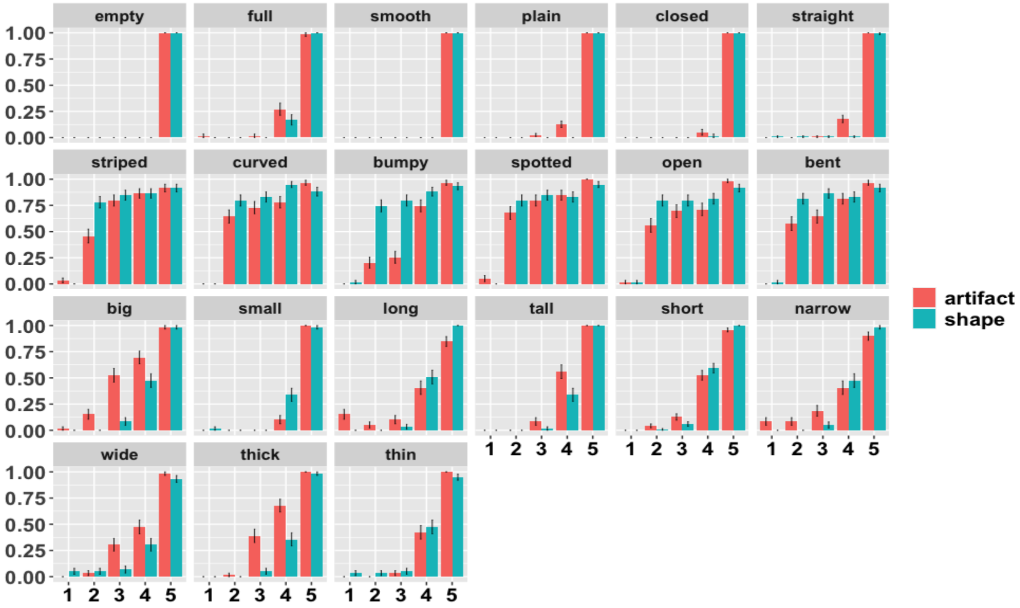


Figure 6: Experiment 2 By-adjective TVJ results. **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

This model revealed a number of effects of interest which are summarized in Table 1. First, the results showed that the distribution of the truth value judgments are different for the three adjective classes. Compared to the relative adjectives, overall there are fewer positive responses for the maximum adjectives and more positive responses for minimum adjectives, although the evidence for the maximum adjective was weaker (since the 95% credible interval did not totally exclude zero). These effects are not surprising, given the consensus in the semantics literature that the three classes of adjectives involve different thresholds (independent of how one derives such differences). It is a little surprising though that there is no clear evidence for an interaction between adjective classes and scale position. Future replication of the current study would be useful to verify whether an interaction could be detected.

Another finding from this model, also shown in Table 1, is that there is an effect of image type, and image type also interacts with different classes of adjectives differently. To better understand the effect of image type on each adjective class, we carried out analyses for each adjective class separately. For each adjective class, the model included the fixed effects of image type, scale position and their interaction. The model also included by-participant, by-image_set, and by-adjective random intercepts and random slopes for the effect of scale position; in addition there was

Effects	Estimate	SE	Lower 95% CrI	Higher 95% CrI
Intercept	-1.49	0.93	-3.23	0.43
Maximum Adj	-2.49	1.28	-4.96	0.06
Minimum Adj	2.62	1.21	0.23	4.95
Scale Position	2.52	0.63	1.23	3.68
Maximum x Scale position	0.98	0.86	-0.77	2.64
Minimum x Scale position	1.43	0.82	-0.11	3.07
Image type	0.73	0.34	0.06	1.39
ImageType x Maximum	0.84	0.58	-0.29	1.99
ImageType x Minimum	-1.99	0.54	-3.05	-0.94

Table 1: Experiment 2: Posterior mean, standard error, 95% credible interval for each effect of interest.

also a by-adjective random slope for the effect of image type. Image type was treatment coded with artifact as the baseline level, and scale position was again treated as a continuous variable and was centered before entering into the model. The results from these models are presented in Table 2. For the maximum adjectives, there is no evidence for an effect of image type. This is likely driven by the fact that there are very few data points for the non-maximum scale position: participants rarely gave a positive response before scale position 5. For the minimum adjectives, there is evidence that participants gave more positive responses for the shape objects than for the artifact objects. For the relative adjectives, there is no clear evidence for an effect of image type.

2.3 Experiment 3: Posterior degrees

2.3.1 Methods

Our final experiment collected empirical estimates of degree posteriors: the probability distribution over the degrees to which language users believe an object has a particular scalar property, after hearing an utterance that describes that object with a gradable adjective.

Participants The experiment was conducted on IbexFarm. As in Experiment 2, shape images and artifact images were tested in two different groups. Sixty-seven participants were recruited for the shape group (mean age: 35; 23 females), and a

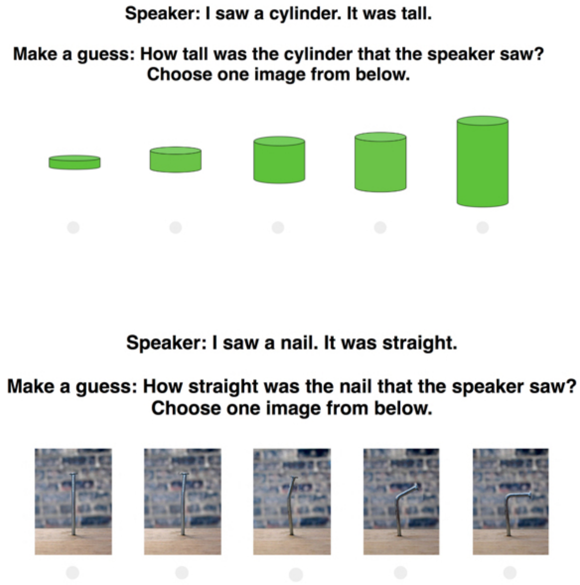
Maximum adjectives	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	0.37	1.14	-1.88	2.60
Image type (shape)	-1.20	1.30	-3.74	1.44
Scale position	1.21	1.30	-1.43	3.76
Image type x Scale position	1.06	1.05	-0.96	3.19
Minimum adjectives				
Intercept	1.62	0.74	0.02	2.95
Image type (shape)	2.54	0.97	0.54	4.38
Scale position	3.38	0.41	2.59	4.16
Image type x Scale position	0.69	0.51	-0.28	1.70
Relative adjectives				
Intercept	-1.78	0.93	-3.42	0.22
Image type (shape)	-1.17	0.80	-2.70	0.45
Scale position	2.29	0.74	0.75	3.66
Image type x Scale position	0.95	0.55	-0.12	2.04

Table 2: Experiment 2: For each adjective class, Posterior mean, standard error, 95% credible interval for each effect of interest.

different group of sixty-eight participants for the artifact group (mean age: 37; 27 females).

Materials The image sets and adjectives used for this experiment, as well as the procedure to pair together the images and adjectives and to distribute them among participants, were identical to Experiment 2.

Procedure The task was set up as follows. A speaker made an utterance in which they first described a visual experience involving a particular type of object (*“I saw an X.”*), and then characterized the object as having a particular scalar property using the positive form of a gradable adjective (*“It was ADJ.”*) Participants were then asked to make a guess about the degree to which the object the speaker mentioned manifests the relevant property (*“Make a guess: how ADJ was the X that the speaker saw?”*) by selecting exactly one of the five objects from the image sets used in Experiments 1 and 2. Example stimuli are shown in Figure 7.



Speaker: I saw a cylinder. It was tall.

Make a guess: How tall was the cylinder that the speaker saw?
Choose one image from below.

Speaker: I saw a nail. It was straight.

Make a guess: How straight was the nail that the speaker saw?
Choose one image from below.

Figure 7: Sample stimuli for Experiment3: Estimating degree posteriors

2.3.2 Results and discussion

The average percentages of choices for each scale point are presented in Figure 8. The results for each individual adjective are presented in Figure 9. Upon visual inspection, the averaged results show the following qualitative patterns: for maximum adjectives, participants consistently chose the maximum degree; for minimum adjectives, the choices distributed among all the non-minimal degrees; and for relative adjectives, the choices were clustered mainly on degrees above the mid-point. The statistical analysis procedure was identical to Experiment 2. We again used the R package “brms” to fit a Bayesian mixed effects logistic regression model predicting people’s choice of an image over not choosing it. But since we do not have a priori predictions as to how participants should update their posterior degree judgments, the statistical results reported below are only for descriptive purpose. Table 3 shows the results when all the data are considered together, and Table 4 shows the results from analyses performed for each adjective class.[7]

[7] One potential question about Experiment 3 is whether participants indeed made their choices based on the speaker’s utterance (i.e. the adjective), or they simply made their choice based on some sort of salient feature of the objects. One way to address this is to compare the results from Experiment 3 with the results from Experiment 1, the elicitation of priors. If the results from Experiment 3 reflect listeners’ belief update based on the speaker’s utterance, their judgments about individual adjectives in this task should be different from their judgments in the prior elicitation task, which did not involve

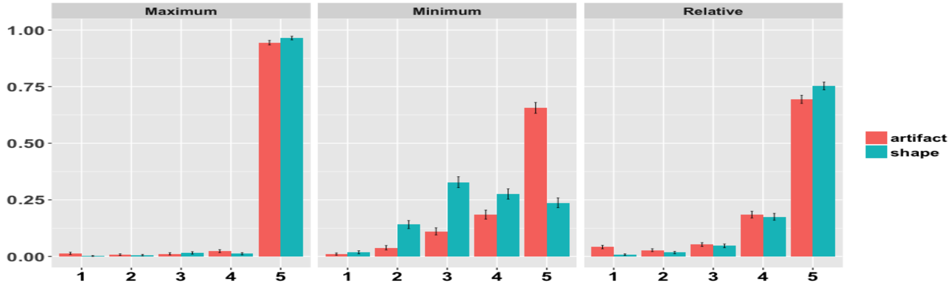


Figure 8: Experiment 3: Percentage of item selection at each scale position for each adjective class.

Effects	Estimate	SE	Lower 95% CrI	Higher 95% CrI
Intercept	-3.76	0.35	-4.44	-3.07
Maximum Adj	-3.05	0.55	-4.09	-1.95
Minimum Adj	1.29	0.48	0.35	2.23
Scale Position	2.38	0.28	1.82	2.91
Maximum x Scale	2.16	0.42	1.30	2.96
Minimum x Scale	-1.13	0.38	-1.88	-0.38
Image type	0.16	0.25	-0.34	0.64
ImageType x Maximum	0.58	0.36	-0.10	1.28
ImageType x Minimum	-0.88	0.30	-1.47	-0.31

Table 3: Experiment 3: Posterior mean, standard error, 95% credible interval for each effect of interest.

3 Model predictions

In this section we present the quantitative predictions of the LG and QF models for degree posteriors and truth value judgments.[8] The outputs of these two models include direct predictions about posterior degrees that can be compared to the empirical estimates of posterior degrees collected in Experiment 3, which we do in Section 3.1. However, these two models do not make direct predictions about truth value judgments, and instead must be supplemented with additional linking

a communicative context. And indeed, visual inspection of the individual adjective results from Experiment 1 (Figure 3) and Experiment 3 (Figure 9) show a clear difference in judgments, indicating that the results of Experiment 3 reflect a posterior update conditioned on the speaker's utterance.

[8] Our technical implementations of these models were adapted from Qing & Franke 2014; we are very grateful to Ciyang Qing for generously sharing his R code with us for this purpose.

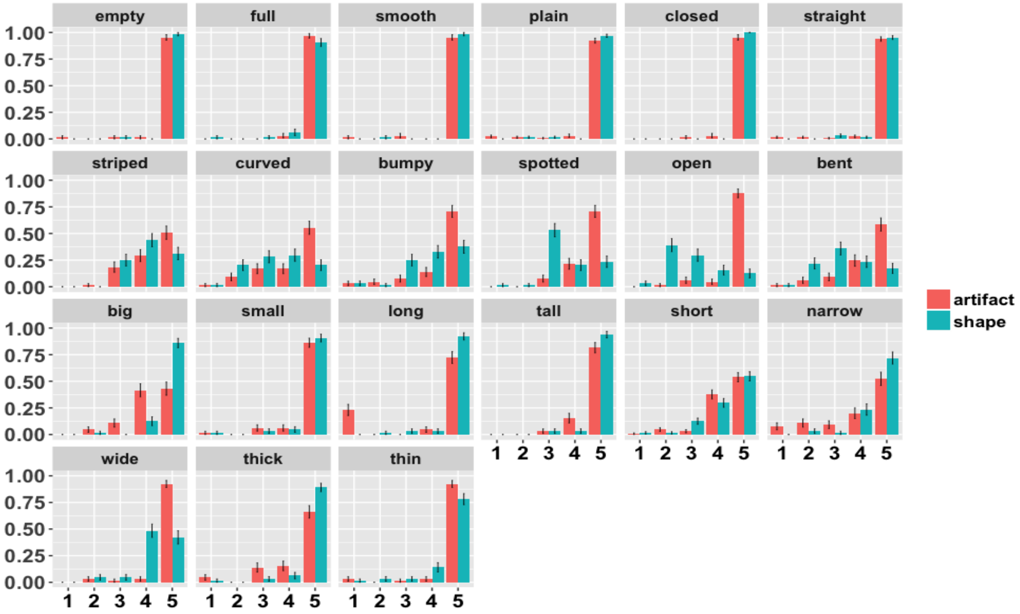


Figure 9: Experiment 3: by-adjective posterior degree judgments. **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

Maximum adjectives	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	-0.14	1.12	-2.37	2.11
Image type	-0.96	1.12	-3.08	1.29
Scale position	0.90	1.45	-1.90	3.84
Image type x Scale position	0.99	0.80	-0.57	2.58
Minimum adjectives				
Intercept	-2.66	0.23	-3.12	-2.23
Image type	0.99	0.28	0.45	1.54
Scale position	1.58	0.20	1.19	1.98
Image type x Scale position	-1.08	0.25	-1.56	-0.59
Relative adjectives				
Intercept	-2.85	0.62	-3.89	-1.46
Image type	-0.77	0.54	-1.82	0.30
Scale position	1.62	0.50	0.49	2.47
Image type x Scale position	0.61	0.43	-0.25	1.45

Table 4: Experiment 3 Posterior Degree: For each adjective class, Posterior mean, standard error, 95% credible interval for each effect of interest

hypotheses about how to relate their outputs to such judgments. We consider two such hypotheses in Section 3.2. The first linking hypothesis is rooted in the truth conditions of gradable adjectives, and the second linking hypothesis is based on pragmatic reasoning about speaker production probabilities.

Since we examined discrete (5-point) scale points in the current study instead of continuous scales, we derived the model predictions for posterior degrees for each scale point for each item we tested in the experiments. An *item* is defined as an image-set (e.g., the set of 5 candles in Figure 1) paired with an adjective. As mentioned in the materials section in Experiment 2 and 3, there are a total of 96 items, 48 for artifacts and 48 for shapes. We likewise derived the model predictions for the threshold distribution for each of the 96 items, also distributed over the five points on the relevant scale. For example, given an utterance *"It is tall"* and an image of five candles with different heights, we generated predictions for the LG and QF models about the probability that a hypothetical listener, upon hearing this utterance, would believe the candle to be d-tall, for d equal to one of the five degrees, and the probability that this listener would think d is the threshold, for d equal to one of the five degrees. For the utterance *"It is short"* and the same image of five candles, we likewise generated its posterior degree distribution and its theta distribution. The by-item predictions of the two models are then correlated with the by-item empirical results from Experiment 2 and 3, with the model predicted result from a particular scale position in a specific item correlated with the empirical result from the same scale position in the same item. This by-item correlation serves as the basis for most of the quantitative model assessment results we present below. But for visualization purpose only, the bar plots we present below are averages of items based on adjective classes, such that items whose adjective belongs to the same class were averaged together. In the Supplementary Material (S4), we also present the by-adjective results of model predictions (in bar plots).

The basic procedure for running the models was as follows. In Experiment 1, we obtained the empirical estimates for the relative frequencies of the choice of each scale position within each image-set (48 image sets total). We use this empirical estimate as a proxy for the prior distribution of degrees. This served as the input to the LG and QF models. But before the priors were entered into the models, for we added 0.00001 to the values of those scale positions that had a relative frequency of 0 (about 5% of the total data) and subtracted 0.00001 from the scale position on the same image-set that had the largest frequency value, since the implementation of the belief update process is not mathematically possible for zero probabilities. As mentioned earlier, each image set was used together with a pair of antonyms. For each member of the antonym pair, their priors are therefore mirror images of each other. For example, the candle set is used for both the trial *"It is tall"* and *"It is short"*, the priors used for these two items came from the same image-set, but are reversed in order. The LG model also requires priors for the threshold θ, which we set to a uniform distribution, as in Lassiter & Goodman (2013). As described in Section 1.2.2 (equation (9)), both models also make use of two free parameters, λ and *cost*. We first present results based on λ=3 and *cost* = 2, which are comparable to the values used by Lassiter & Goodman (2013) and Qing & Franke (2014). We will discuss a more systematic parameter selection process in section 3.2, and also provide a quantitative model fitting there.

3.1 Posterior degrees

Figure 10 shows the LG and QF model predictions for posterior degrees, alongside the empirical posteriors collected in Experiment 3, for each adjective class. Visual inspection of these figures suggests that both models make good predictions, with the overall patterns of the model predictions similar to the actual patterns in the empirical data. The aggregated patterns in Figure 10 may mask more fine-grained by-item variations, and we will present a quantitative model assessment at the item level in the next section. Qualitatively speaking at least, the model predictions captured two key features present in the empirical data. First, the models correctly predicted the general differences between the three classes of adjectives. For maximum absolute adjectives, the maximum degree point has the highest probability, by and large excluding any other degree points; for minimum adjectives, the probabilities are much more evenly distributed on all degrees above the scale point one, peaking around the middle part of the scale; and for relative adjectives, although the maximum degree also has the highest probability, other scale positions especially scale point 4 also has some amount of probability mass. Second, the model predictions also captured the general difference between the artifact and shape objects, in keeping with the qualitative patterns in the empirical results.

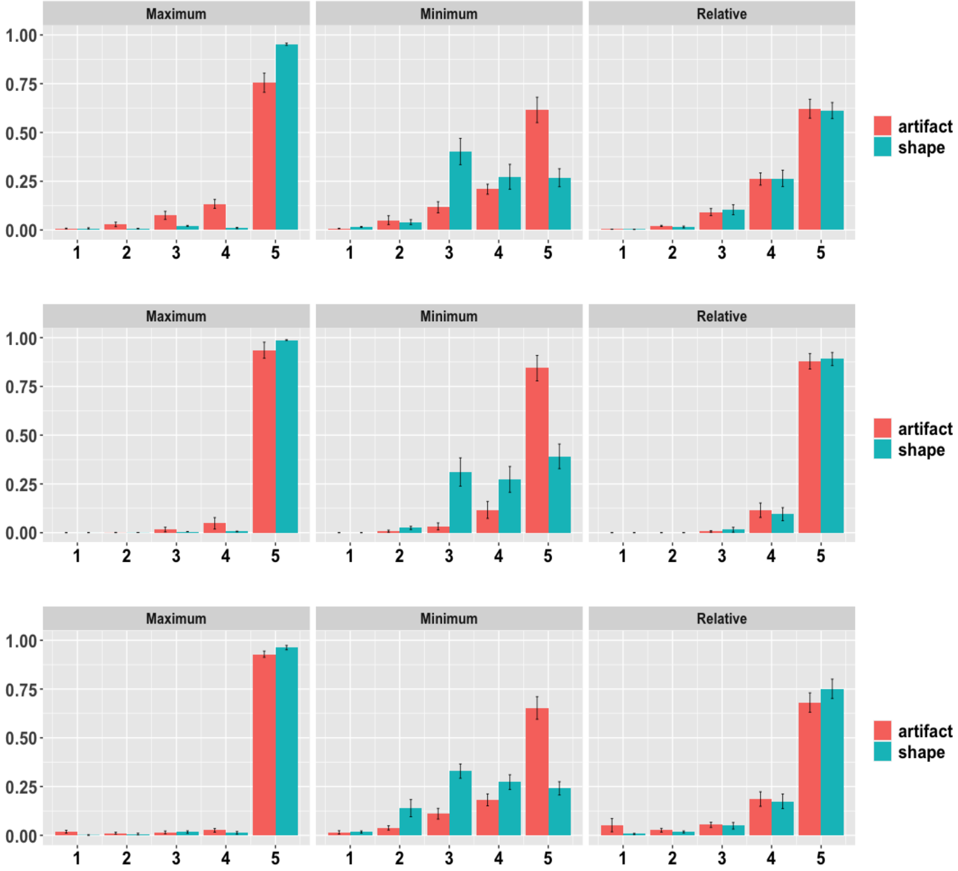


Figure 10: Comparison of LG predictions (top), QF predictions (middle) and Experiment 3 empirical estimates (bottom) of posterior degrees.

Overall, the posterior degree predictions that we derived using empirical priors provide support for the general hypothesis guiding Lassiter & Goodman (2013) and Qing & Franke (2014): that the information communicated by utterances involving gradable adjectives is impacted by prior beliefs about how objects distribute along a scalar dimension. Our results show that the model predictions track human judgments, both about differences based on adjective class and on differences based on the type of object being talked about.

3.2 Truth value judgments

The LG and QF models do not make direct predictions about truth value judgments. To derive their model predictions for truth value judgments, we consider two possible linking hypotheses.

3.2.1 Linking truth judgements to semantic content

Our first hypothesis about how to link Bayesian models of communication with gradable adjectives to judgments about truth is that such judgments are grounded in knowledge of semantic content, specifically in knowledge of the threshold-based truth conditions of gradable adjectives. In particular, we hypothesize that upon hearing an utterance such as *"That is tall"*, used to describe a particular object, an individual's judgment about whether the utterance is true is a function of their belief about the likelihood that the object's height is at least as great as what they take the threshold to be. That is, given the utterance, the more likely the listener takes it to be that the object's height is at least as great as the threshold, the more likely they are to judge the utterance as true.

Our implementation of this hypothesis takes advantage of the fact that, in addition to making predictions about posterior degree distributions, the LF and QF models also make predictions about (posterior) threshold distributions. We may then derive model predictions for truth judgments according to the equation in (12): the probability that an utterance will be judged true of an object at scale position *i* is the probability that the threshold falls on a position below or equal to position *i*. In the current case, since we only consider discrete scale positions, the target probability is equal to the sum of posterior threshold probabilities at each scale position $j \leq i$.

(12) $P(\text{“}x_i\ is\ adj\text{”}\ \text{is TRUE}) = P(\theta \leq d_i) = \sum_{j \leq i} P(\theta_j)$

Figure 11 shows the model-predicted threshold predictions of the LG and QF models for each adjective class (λ=3 and *cost* = 2, as in section 3.1), and Figure 12 shows the corresponding model predictions for truth value judgments based on the linking hypothesis in (12), as well as the empirical truth value judgements we collected in Experiment 2. A qualitative comparison between the model predictions for truth value judgments and the empirical results indicates a non-satisfactory match. For relative adjectives, both models generate predictions that are, by and large, consistent with the empirical data, but the model predictions for absolute adjectives are not. For maximum adjectives, the LG predictions fail to capture their endpoint-oriented truth conditions. The QF model does better at locating most of the probability mass for truth value judgments at the upper end of the scale, but it predicts that artifacts should be treated more categorically than shapes, which is the opposite of the empirical results. For the minimum adjectives, the LG predictions are more or less consistent with the empirical findings: all degree points other than the lowest scale point receive a substantial amount of “true" judgments, and the shape objects have a higher percentage of acceptance in the middle range of the scale; although the model predictions seem more gradient than the empirical results. The QF predictions for the minimum adjectives, on the other hand, are less good, failing to generate a sufficient amount of acceptance for degree points in the middle range of the scale for both shapes and artifacts.

Descriptively speaking, then, it appears that both the LG and the QF models fall short of accurately predicting human truth value judgments for absolute adjectives, given a linking hypothesis based on truth conditions. The LG model does better with minimum adjectives, but it treats the maximum adjectives as if they were relative. The QF model does the opposite: it performs fairly well on the maximum adjectives, though it fails to capture the shape/artifact distinction, but it treats the minimum adjectives more like the maximum ones.

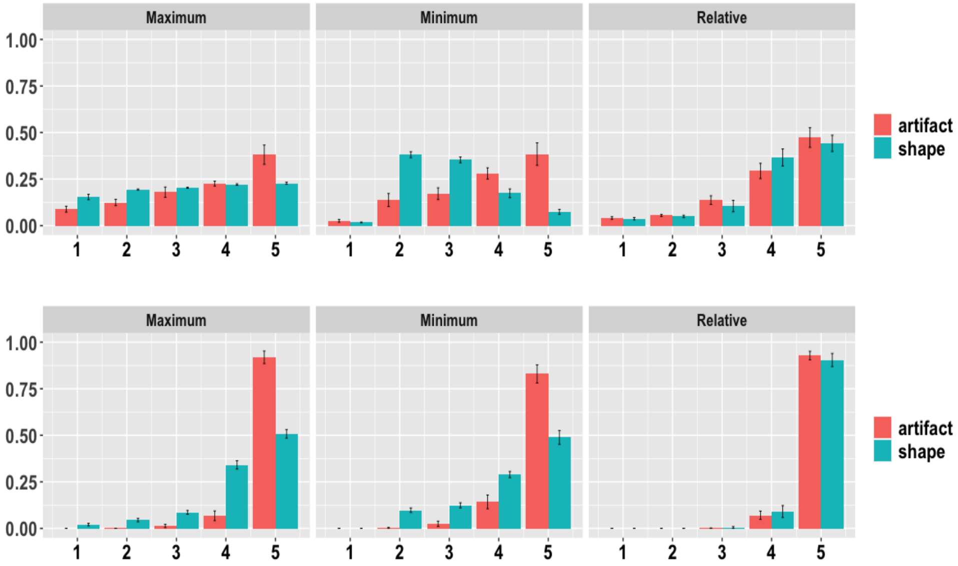


Figure 11: LG (top) and QF (bottom) predictions for thresholds.

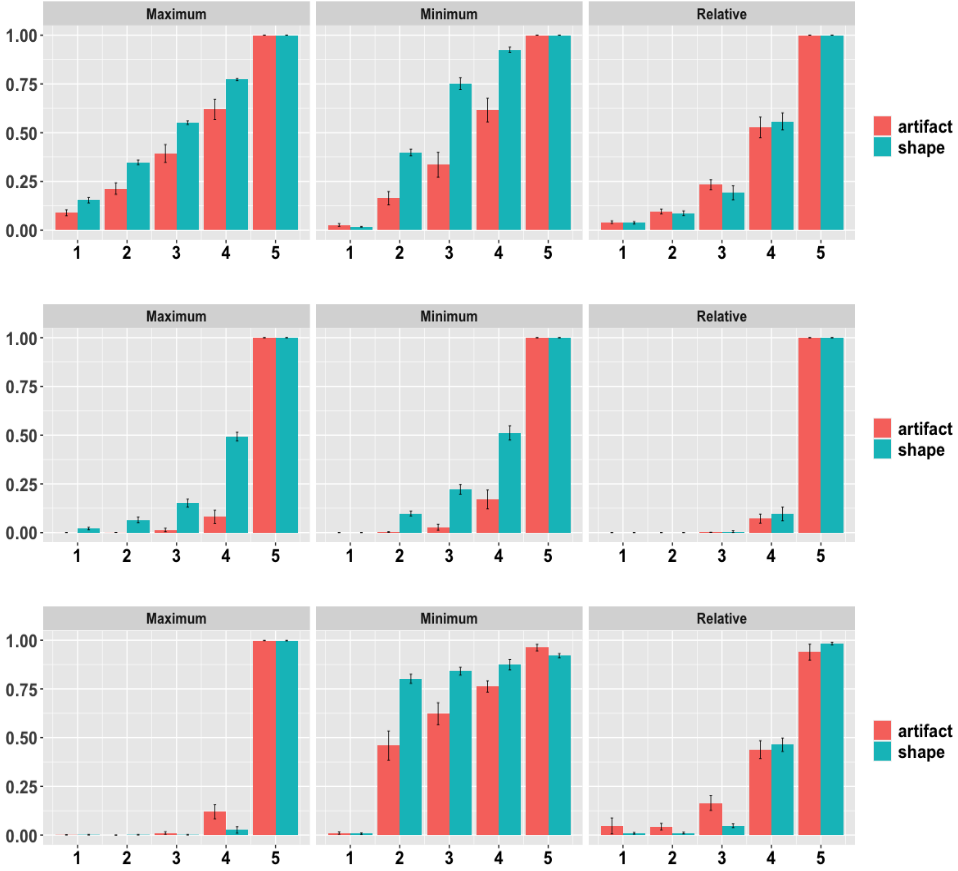


Figure 12: Comparison of LG predictions (top), QF predictions (middle) and Experiment 2 empirical judgments (bottom) of truth values.

One potential explanation for the models' non-ideal performance on threshold predictions and truth value judgments is that it reflects a problem with our methodology for deriving empirical degree priors. If this were the case, however, then we should have also seen weak model performance on degree posteriors, but this was not the case.

A second potential explanation for the models' weakness is that we were not using optimal values for the free parameters λ and *cost*. To test this possibility, we formed 110 different (λ, *cost*) settings by fixing λ to an integer value between 1 and 10 and cost between 0 and 10, and we obtained the LG and the QF model predictions for the truth value judgments and the posterior degree probabilities using each setting. For each parameter combination of λ and cost, the model predictions for each item were correlated with the by-item experimental data from Experiment 2 and 3 to derive a R^2 score, which we used to select the best parameter values. Across the R^2 scores obtained from the 110 (λ, *cost*) combinations, for the LG model, the R^2 values range between 0.26 and 0.81 for the posterior degree judgments, and between 0.36 and 0.61 for the truth value judgments; and for the QF model, the R^2 values are between 0.25 and 0.82 for the posterior degrees, and between 0.52 and 0.63 for the truth value judgments. Among all the models, the best ones in general have a better R^2 value for the posterior degree judgments than for the truth value judgments. Importantly, the parameter settings that we described above, with $\lambda = 3$ and *cost* = 2, was among the choices that showed the best R^2 scores.

Using the parameter values $\lambda = 3$ and *cost* = 2, we further correlated the model prediction and the experimental results for each adjective class. The correlation results are shown in Table 5 and Figure 13. When all items are considered, both the LG and the QF models make better predictions for posterior degree predictions than for truth value judgments, in line with the qualitative conclusion we drew earlier. But we also note that for the posterior degree predictions, both models performed relatively poorly on minimum adjectives and we will come back to this in later sections. We can also see that the models differ in where they go wrong on truth values: the LG model shows a weaker performance with maximum adjectives, while the QF model is weaker with minimum adjectives.

	LG (R^2)		QF (R^2)	
	TVJ	POSTERIOR	TVJ	POSTERIOR
All items	0.6	0.78	0.6	0.82
Maximum	0.57	0.94	0.83	0.97
Minimum	0.68	0.55	0.38	0.58
Relative	0.82	0.69	0.76	0.78

Table 5: With $\lambda = 3$ and *cost* = 2, by-item correlations between the LG and QF model predictions and the experimental results

3.2.2 Linking truth judgments to speaker production probability

The linking hypothesis we adopted in Section 3.2.1 was grounded in the idea that human truth judgments are based on knowledge of semantic content, specifically, in the current case, the threshold-based truth conditions of gradable adjectives. One problem for this idea, however, is that truth value judgment tasks may not be driven exclusively by individuals' knowledge of truth conditions. Truth judgments are often noisy, they can be influenced by experimental manipulations, and they may be based on information that goes beyond semantic content strictly speaking. (See e.g., the

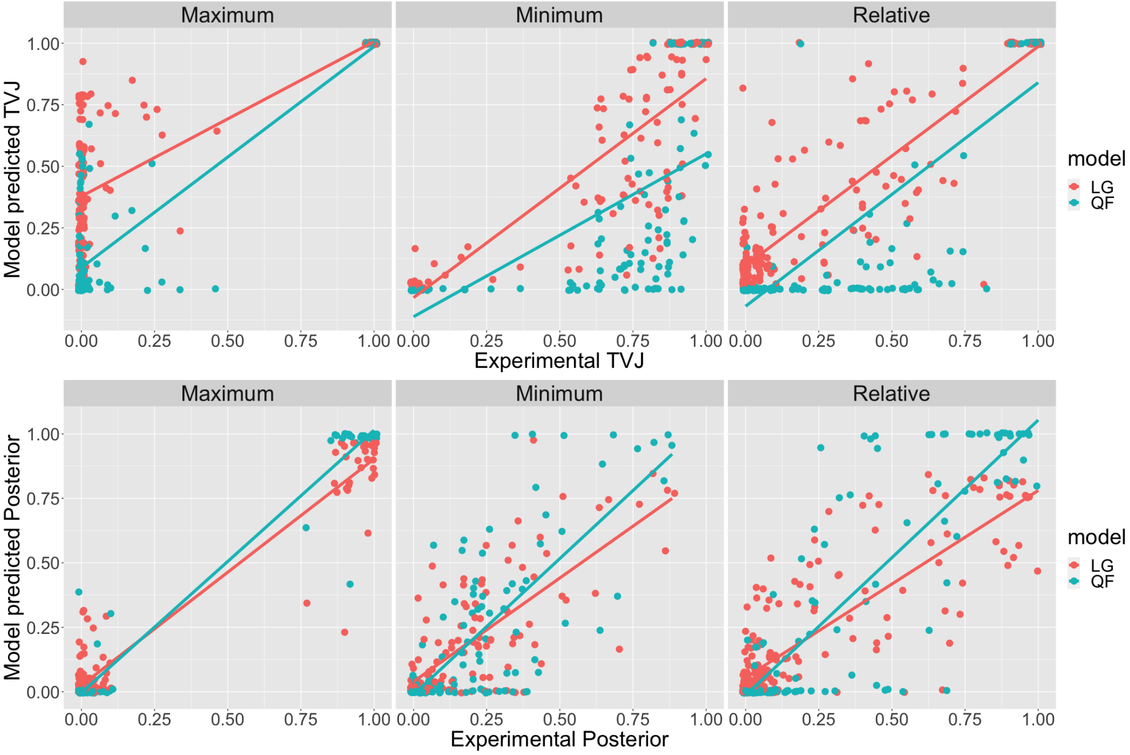


Figure 13: By-item correlation between model predictions and experimental results. Top:Truth value judgments; Bottom: Posterior degree judgments. X-axis denotes experimental results and Y-axis denotes model predictions.

use of truth value judgment tasks in experimental work on scalar inferences, such as Noveck 2001, Papafragou & Musolino 2003, Doran et al. 2012, van Tiel et al. 2016, Ronai & Xiang 2020.)

As an alternative to a linking hypothesis based on truth conditions, several recent studies within the RSA framework have proposed to model truth judgments in terms of the probability that a speaker would produce a particular utterance (Tessler & Goodman 2019, Jasbi et al. 2019, Waldon & Degen 2020). The general idea is that there is a close relationship between a pragmatic listener's judgment of the truth of a sentence S (in a particular context) and a cooperative pragmatic speaker's decision about whether to utter S (in that context), since a crucial factor that a pragmatic speaker takes into consideration when deciding whether to utter a sentence is how a pragmatic listener would interpret it.

Recall that the RSA model has the flexibility to model the recursive reasoning between a speaker and a listener. The current implementation only included three levels: a literal listener L_0, a speaker S_1 that reasons about the literal listener, and a pragmatic listener L_1. In order to determine whether truth value judgments of a pragmatic listener could be modeled by speaker probabilities, we need to extend the current RSA model implementation to the next level pragmatic speaker S_2 (i.e. add another level of recursive reasoning on the current RSA model implementation). The S_2 pragmatic speaker makes its production decisions by considering the interpretation outcome of the previous level pragmatic listener L_1, and the production-based linking hypothesis would predict a close correspondence between the model predicted production probabilities of S_2 and the human truth value judgments in Experiment 2. The details of implementing the production model are described in the Supplementary Materials (S3). Using the production model, we generated model predictions of the probability of a pragmatic speaker choosing to use the relevant adjective for each scale position in each of the 96 items used in Experiment 2. Figure 14 shows the model-predicted production probabilities aggregated by adjective class (top), compared to the human truth value judgments from Experiment 2 (bottom).

When we correlated the model predicted by-item production probability with the by-item truth value judgment results collected in Experiment 2, we obtained a $R^2 = 0.67$ when all the items were considered, showing a decent overall performance of the model. However, a closer look reveals some weaknesses. As Figure 14 and Figure 15 indicate, production probability matches human truth judgments for relative adjectives relatively well ($R^2 = 0.72$), but it has weaker performance in predicting human truth judgments for absolute maximum adjectives ($R^2 = 0.53$). In particular, human judgments consistently target the maximal degree for the maximum adjectives, but this is not the case for production probability. Figure 14 also suggests that the production probability model did not seem to capture the artifact vs. shape

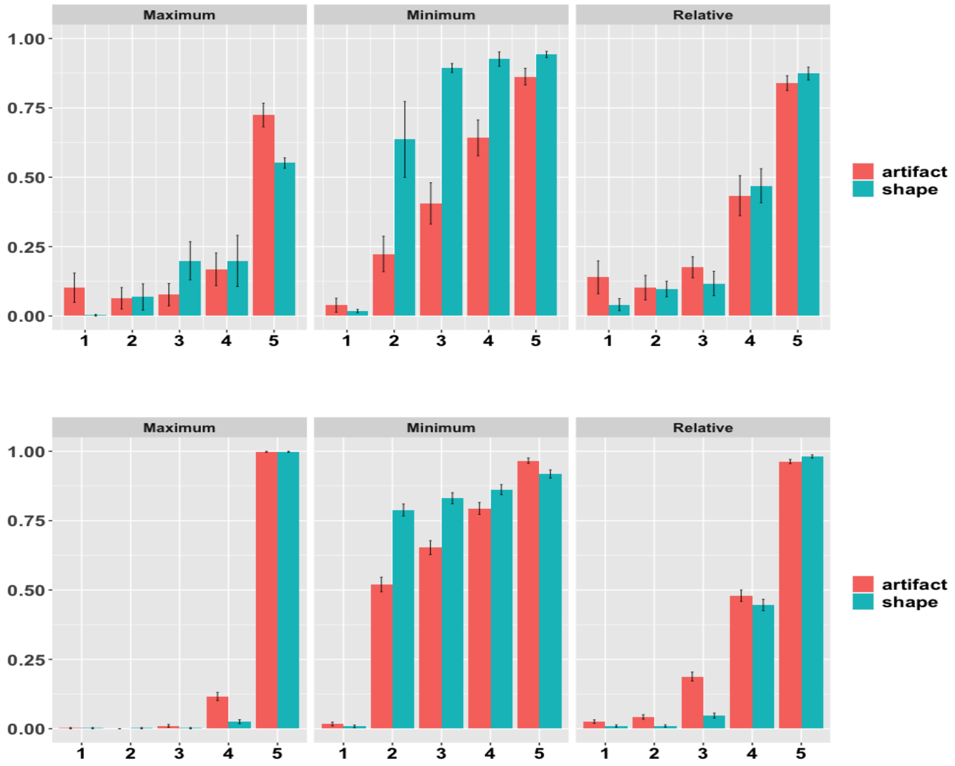


Figure 14: Model predicted production probability of a pragmatic speaker (top) and empirical truth value judgments from Experiment 2 (bottom).

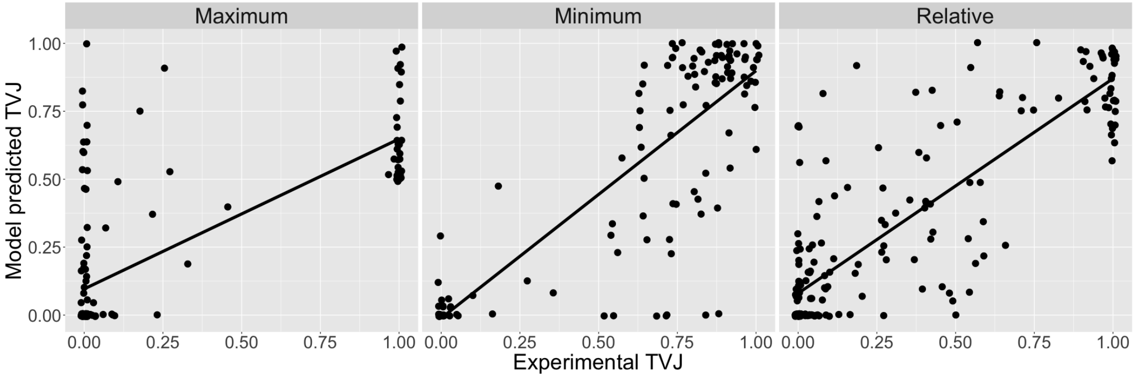


Figure 15: By-item correlation between model predicted production probability (y-axis) and experimental TVJ results (x-axis)

distinction for maximum adjectives. In the case of absolute minimum adjectives, production probability showed decent performance overall ($R^2 = 0.67$).

Overall, then, modeling truth judgments as speaker production probability can capture some of the human data but not all of it, and modeling truth judgments in this way does not improve on the results of the truth conditional linking hypothesis that we considered in Section 3.2.1 (compare the R^2 scores above to the ones for the LG TVJ predictions in Table 5). Empirically speaking, for the current case at least, the weaknesses of the two linking hypotheses about truth value judgments are actually similar (e.g., both performed the worst for the TVJ judgments of maximum adjectives). From a more general theoretical perspective though, a production-based linking hypothesis removes a listener's interpretation of an utterance as the main driving force of their truth value judgments; instead, a listener's truth value judgment is almost solely based on how likely they think a speaker would utter the target sentence. A cautionary note is that a key theoretical assumption underling the production-based linking hypothesis — that there is a close (mirroring) relation between comprehension and production — is an issue under active debate in the larger context of language processing. There are still many open questions about the extent to which production and comprehension processes are related to each other and how the interactions can be implemented in mechanistic ways (Ferreira 2019, Pickering & Garrod 2013).

3.3 Summary

In sum, Bayesian pragmatic models appear to be good at predicting posterior degree distributions based on empirical priors, but less good at predicting the truth value judgments. Their weaker performance on truth value judgments holds across two different linking hypotheses that connect the model predictions to human judgments. Before we discuss the implications of these findings, we note that our conclusions about the model predictions are only as good as the choices we made in conducting the experiments and implementing the models. There are a number of open issues that could be improved in future work.

On the experimental side, as we have discussed earlier, more work is needed to develop paradigms that can accurately elicit different types of prior beliefs of language users. Another potential concern with the current study is that we only created five-point scales, which could be too coarse-grained to capture the real-world patterns. Future work could use more dense scales. On the modeling side, the question about how to link model predictions to truth value judgment concerns not just the current case study; given how the truth value judgment task is widely applied in semantics and pragmatics research, a deeper understanding of how to model such behavioral judgments is critical for our theory building. Another question is how much of the model's limitations are due to technical implementations that could be improved. For example, one reviewer suggested that the softmax function in the model implementation may have contributed in part to some of the limitations of the QF model predictions. Lacking of a plausible way to substitute this function in the current implementation of the models, we will have to leave this as a question for future work.

4 General discussion

We have seen that Bayesian pragmatic models can capture several important features of human judgments about utterances involving gradable adjectives. The models make good predictions, qualitatively and quantitatively, about posterior degree judgments, and they also capture differences in these judgments that correlate with the types of objects that the judgments are about (artifacts vs. shapes). The weakness of the models mainly resides in capturing truth value judgments, especially those involving absolute adjectives: the LG model is weaker for maximum adjectives, and the QF model is weaker for minimum adjectives. Ideally, one would like a model that can account for both the posterior degree judgments and the truth value judgments, since both seem to reflect what a gradable adjective *means* to people. Given that the Bayesian pragmatic approaches reflect the state of the art on the former, a question then is whether we can improve their performance on the latter by supplementing them with additional features. We suggest that one way to do this is to integrate the threshold conventions posited by semantic theory with the Bayesian pragmatic reasoning.

Focusing on the hypothesis that truth judgments are based on truth conditions, the models' weaknesses can be traced to the model's predictions of the threshold distribution. In other words, it is the predicted threshold distributions in Figure 11 that drive the (inaccurate) predictions about truth value judgments in Figure 12. This leads to the following question: what kind of threshold distributions would more closely derive empirical truth value judgments? We did not design an experiment to directly elicit participants' judgments about where they believed thresholds were located (and it is not clear that such an experiment would be practical), but we can use subjects' truth value judgments in Experiment 2 (see the bottom panel of Figure 12) and our hypothesized link between thresholds and truth value judgments (the equation in (12)) to compute a threshold distribution from the empirical truth value responses. Specifically, for the five scale positions, for scale position 2 and above, the probability of a given scale position to be chosen as the threshold is computed by taking the averaged truth value response at that scale position, and subtracting the truth value response from the previous scale position. For scale position 1, we simply took the truth value response at that position, without subtracting anything.

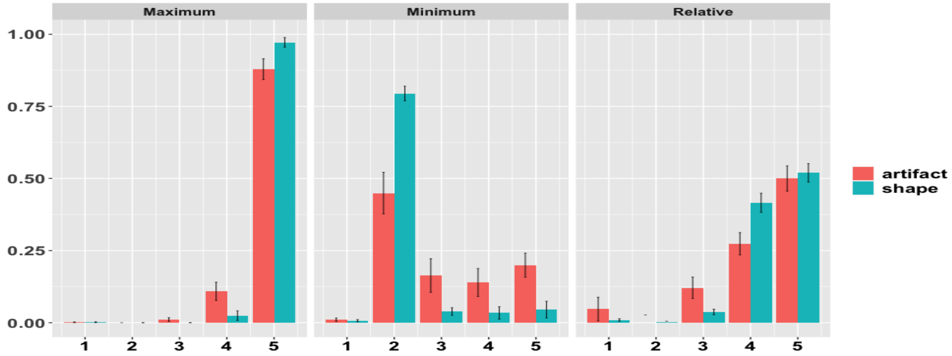


Figure 16: Threshold distribution based on empirical truth value judgments in Experiment 2, at the group level for each adjective class.

The result of this computation, at the group level for each adjective class, is shown in Figure 16.

The "reverse engineered" threshold distribution in Figure 16 is very close to what we would expect from theories that postulate semantic conventions for associating adjectives that use closed scales with endpoint-oriented thresholds. Maximum adjectives have thresholds that are largely aligned with the maximum degree on the scale, nearly perfectly so for shapes, and with some relaxation for artifacts. For minimum adjectives, scale point 2 is the most likely threshold, with shapes again having a stronger preference for a strict endpoint orientation than artifacts. Relative adjectives show a more gradual pattern of threshold distribution, but here too, shapes show a more categorical distribution than those for artifacts. The observed shape vs. artifacts difference has some important theoretical implications, and we will come back to it later. But for the moment let us gloss over this difference and refer to the reverse engineered thresholds as "semantic thresholds."

As we noted in Section 1.3, there are two kinds of questions one can ask about the meanings of utterances involving positive form gradable adjectives: (S) What are the truth conditions of such utterances? And (P), what information do such utterances communicate? Given that the current results seem to show that the semantic theories (and their corresponding threshold conventions) provide a good answer to (S) while the Bayesian pragmatic theories provide good answers to (P), we can ask whether a model that incorporates the semantic and Bayesian approaches can do a better job of capturing the full pattern of human responses.

To answer this question, we computed two additional models. First, we created a "semantic threshold model" by replacing the original threshold distribution $Pr(\theta)$ used in the QF model — which, recall the discussion in section 1.2.2, is considered to represent the optimal thresholds for best communicative efficiency given people's knowledge of the priors — with a new distribution $Pr_{sem}(\theta)$ that we calculated for each item in the same way that we "reverse engineered" the group-level semantic threshold distribution in Figure 16 (i.e., based on the linking hypothesis in (12)). The original QF model is shown in (QF) below (repeated from equation (11), and also see the model details there). Our semantic threshold model is shown in (ST).[9]

(QF) $P_L(d \mid u) \propto P_S(u \mid d, Pr(\theta)) \times P_L(d)$

(ST) $P_L(d \mid u) \propto P_S(u \mid d, Pr_{sem}(\theta)) \times P_L(d)$

Second, we linearly combined the (original) QF model with the ST model to create a "hybrid" model, which implements the hypothesis that communication with gradable adjectives involves simultaneous consideration of both pragmatic thresholds derived from a utility based computation (as in QF) and semantic thresholds (as in ST).

(Hybrid) $P_L(d \mid u) = QF \cdot \beta \, + \, ST \cdot (1-\beta)$

Specifically, the hybrid model introduces a free parameter β, such that the degree posterior corresponds to a proportion β of the posterior degree probability computed by the QF model, plus a proportion $(1-\beta)$ of the posterior probability computed by the ST model. The hybrid model thus has three parameters: β, plus the λ and *cost* parameters that factor into the computation of $Pr(\theta)$ in the QF model. We set $\lambda = 3$ and *cost* = 2, in line with the QF model parameters chosen earlier; for β, we searched through values between 0 and 1 at 0.01 increments, settling on 0.62, which yielded one of the highest R^2 scores between the model predicted posterior and the empirical posterior values.[10]

Figure 17 presents the posterior degree probabilities by-adjective-class predicted by the ST and Hybrid models, and Figure 18 shows the by-item correlation between the model predictions and the empirical estimates of posterior degrees collected in Experiment 3, as well as the by-item correlations based on the original QF model (shown earlier in Figure 13). For the ST model, the correlations showed a $R^2 = 0.98$ for maximum adjectives, $R^2 = 0.19$ for minimum adjectives and $R^2 = 0.58$ for relative adjectives. For the hybrid model, we see $R^2 = 0.97$ for maximum adjectives, $R^2 = 0.32$ for minimum adjectives and $R^2 = 0.8$ for relative adjectives.

At first glance, both models showed the worst performance on minimum adjectives, and worse than the original QF model (see Table 5). However, since the R^2 measures were based on the point estimates of the models' predictions, and so

[9] We used the QF model rather than the LG model for this exercise for the simple reason that it was easier to implement a change in the calculation of the threshold distribution in the former, in which threshold distribution is calculated independently of degree posteriors.

[10] The parameter combination with $\lambda = 3$, *cost* = 2, and $\beta = 0.62$ is also among the best combinations when we searched all three parameters together.

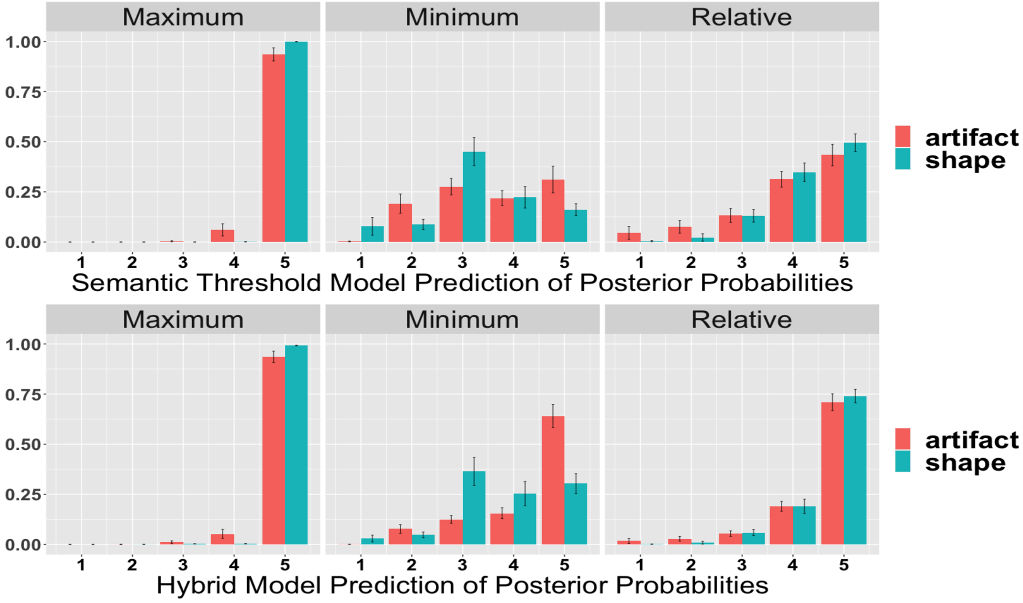


Figure 17: Posterior degree predictions of ST (top) and Hybrid (bottom) models.

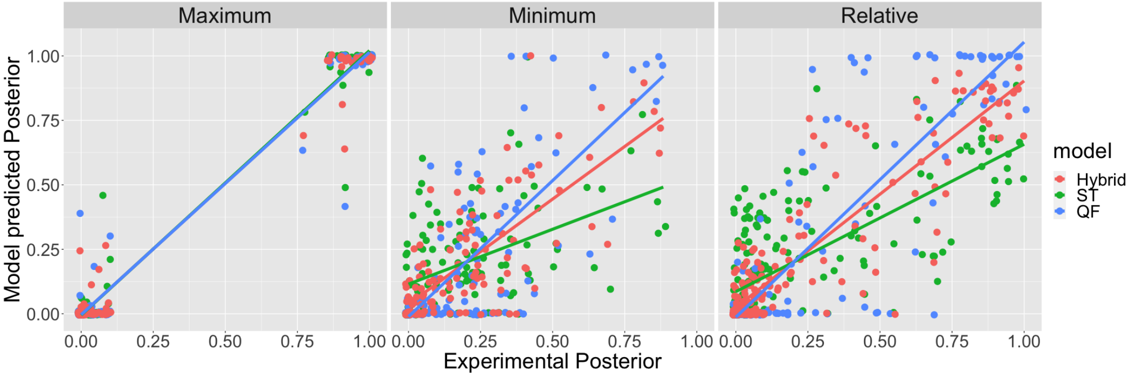


Figure 18: Correlations between experimental posteriors and model predicted posterior probabilities

may not best capture their relative strengths, we further compared them using a likelihood ratio test. This test takes a pair of models and compares the probability of obtaining the observed data based on each of them, thus providing a quantitative assessment as to which is more likely to generate the observed data. We conducted a likelihood ratio test of the relative performance of our three models as follows. First, given the QF, ST and Hybrid models with the respective parameters specified as earlier, we generated the model-predicted by-item posterior probabilities at each of the five scale positions. Treating the probability distribution at each scale position as a normal distribution, we then bootstrapped (using the R function boot()) the mean and the standard deviation of the normal distribution for the ratings at each scale position under each adjective class. After that, we calculated, for each scale position, the probability density of the mean observed posterior rating (rescaled as *z*-scores) given the bootstrapped parameter values. Finally, for each adjective class, we calculated the likelihood of a model by multiplying the probability density of the mean ratings from all five scale positions in that model.

The results of our model comparison are presented in Table 6 as the log-likelihood ratio between each pair: Hybrid vs. ST, Hybrid vs. QF, and QF vs. ST. A value of zero indicates the pair have the same likelihood of generating the data; a deviation from zero means they do not, with larger deviations corresponding to larger differences. A positive value indicates better performance of the first member of the pair; a negative difference indicates better performance of the second member of the pair. The overall results of the model comparison indicate that Hybrid is the best of the three, while QF is better than ST when it comes to maximum adjectives, and ST is better than QF when it comes to minimum and relative adjectives. Although the log likelihood ratio values in Table 6 can help us gauge the relative strength of different models, we caution against interpreting the specific numbers at their face value, since some of these values may have been inflated due to extremely small likelihood values at certain scale positions.[11]

The model comparison exercise serves as a proof of concept that the threshold conventions posited by semantic theory can be combined with Bayesian models to derive the communicative content of gradable adjectives. In particular, the hybrid model that considers both the semantic thresholds and the pragmatic thresholds

[11] For our calculation, we replaced all estimated probability density values smaller than 0.001 (on the rescaled (0,1) standard normal distribution) with the value 0.001. Practically speaking, a probability density smaller than 0.001 indicates an extremely small probability, and making further distinctions between two very small values, such as e-3 and e-5, would only inflate the differences between two models (since the final likelihood of a model is calculated as the product of five probability density values). The relatively large difference between the QF and the other two models for minimum and relative adjectives in Table 6 was largely due to the fact that under the QF model, there were multiple extremely small likelihoods for obtaining the empirical mean, especially over some lower scale positions for the minimum and relative adjectives.

	Hybrid vs. ST	Hybrid vs. QF	QF vs. ST
Maximum	2.0	-0.12	2.12
Minimum	0.88	10.7	-9.78
Relative	1.02	13.1	-12.1

Table 6: Log likelihood ratios between models

performs better than the original QF model in generating the observed data on posterior degree judgments. Together with the result that semantic threshold convention can also naturally deliver human truth value judgment results, at least under the linking assumption in section 3.2.1, a model that integrates semantic threshold conventions with Bayesian pragmatic reasoning would be empirically more preferred over Bayesian reasoning alone. This general conclusion is broadly in line with results from van Tiel et al. (2021), which compared different models on their ability to capture human's production choices of quantifiers. The best model emerged from their study was the one that combined both truth-conditional semantics and probabilistic pragmatic reasoning. The model that contained both a semantic and a pragmatic module outperformed those that only contained a single module.

While our results highlight the importance of thresholds motivated by semantic conventions, they also reinforce earlier work which argues that the traditional semantic accounts need refinement. In particular, the threshold distribution in Figure 16, while distinct from the model-predicted threshold distributions, still shows an influence of empirical priors: the distribution of shape thresholds is more categorical compared to artifact thresholds, which plausibly indicates the influence of prior beliefs about distributional differences between shapes and artifacts. This generalization holds not just for relative adjectives, but also for absolute adjectives that, under an account like the one in Kennedy 2007, would have more categorical endpoint oriented thresholds. A number of recent studies have argued for comparison class or comparison distribution uncertainty in the semantics of gradable adjectives, including absolute adjectives (see e.g. Solt 2009, Foppolo & Panzeri 2011, Toledo & Sassoon 2011, Qing 2020, Bumford & Rett 2021); our results suggest that a theory of thresholds that best reflects human truth judgments will need to involve an integration of semantic convention and pragmatic reasoning. (See Qing 2020 for a recent proposal along these lines.)

Finally, we should note that both the LG/QF models and the traditional semantic account have assumed homogeneity of thresholds within a given adjective class. The LG and QF models assume that all adjective types — relative, maximum absolute and minimum absolute — have meanings that introduce threshold uncertainty; whereas under the traditional semantic account, there is threshold uncertainty for relative adjectives and threshold certainty for absolute adjectives. It is therefore worth noting that Qing (2020) has recently argued that at least minimum absolute adjectives are, in fact, semantically ambiguous between interpretations in which the threshold is uncertain (and valued in the same way that it is for relative adjectives) and interpretations in which it is bound by an existential quantifier, and so is indeterminate but not uncertain in its truth conditions. Whether Qing's arguments are correct, and if so, what further implications they have, should be a focus of future work. But if it is true that there is ambiguity about whether an absolute adjective's threshold could be uncertain, this may potentially provide a way to understand an earlier observation: that the model predictions of posterior degrees showed the worst performance for minimum adjectives. This was true both when the thresholds were pragmatically derived (see Table 5), and when the thresholds consistent with semantic theory were used (see the model outcome above from the ST and Hybrid model). Crucially, for these models, minimum adjectives (and in fact each adjective class as well) were treated uniformly: they either all used pragmatic thresholds or all used semantic thresholds. Even in the hybrid model, the way different thresholds are combined is assumed to be the same across all adjectives within the same adjective class. If minimum standard adjectives are in fact ambiguous in actual usage, then it is possible that our human subjects treated them as having fixed (absolute) thresholds on some trials and variable (relative) thresholds on others (e.g., as suggested by an anonymous reviewer, using fixed thresholds more often for truth judgments and variable thresholds more often for posterior degree estimations), and that the relatively poor model performance is due to the fact that the models treated them homogeneously.

5 Conclusion

Gradable adjectives differ from other context-dependent expressions because the contextual thresholds which fix their extensions in particular contexts of use are uncertain. A full semantic and pragmatic account of such expressions must therefore answer two questions about utterances involving predications of gradable adjectives: what are their truth conditions, and what information about degree do they communicate? In this paper, we have shown that, starting from empirically derived prior distributions over degrees, Bayesian pragmatic models do a better job at capturing human judgments about the degree information that gradable adjectives communicate, something that traditional semantic analyses cannot explain, than capturing human judgments about truth conditions. Supplementing the current Bayesian pragmatic models with a semantic convention for associating close-scale adjectives with endpoint-oriented thresholds makes it possible to simultaneously capture both human judgments about truth conditions and human judgments about what is communicated.

References

Bumford, Dylan & Jessica Rett. 2021. Rationalizing evaluativity. In *Proceedings of sinn und bedeutung*, vol. 25, 187–204.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, Articles* 80(1). 1–28. http://dx.doi.org/10.18637/jss.v080.i01. https://www.jstatsoft.org/v080/i01.

Burnett, Heather. 2016. *Gradability in natural language: Logical and grammatical foundations* Oxford Studies in Semantics and Pragmatics. Oxford, UK: Oxford University Press. http://dx.doi.org/http://dx.doi.org/10.1093/acprof:oso/9780198724797.001.0001.

Chen, Sherry Yong & Bob van Tiel. 2021. Every ambiguity isn't syntactic in nature: Testing the rational speech act model of scope ambiguity. *Proceedings of the Society for Computation in Linguistics* 4(1). 254–263. http://dx.doi.org/https://doi.org/10.7275/h3rp-m711.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 3, de Gruyter. http://dx.doi.org/http://dx.doi.org/10.1515/9783110253382.2297.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb & Rachel E Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 124–154. http://dx.doi.org/http://dx.doi.org/10.1353/lan.2012.0008.

Ferreira, Victor S. 2019. A mechanistic framework for explaining audience design in language production. *Annual review of Psychology* 70. 29–51. http://dx.doi.org/http://dx.doi.org/10.1146/annurev-psych-122216-011653.

Foppolo, Francesca & Francesca Panzeri. 2011. When *straight* means *relatively straight* and *big* means *absolutely big*. Paper presented at the 31st Incontro di Grammatica Generativa, Rome, Italy, February 24-26.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. Hampshire: Palgrave Macmillan. http://dx.doi.org/http://dx.doi.org/10.1057/9780230210752_4.

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

Franke, Michael, Fabian Dablander, Anthea Schöller, Erin Bennett, Judith Degen, Michael Henry Tessler, Justine T Kao & Noah D Goodman. 2016. What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data. In *Cogsci 2016: The annual meeting of the cognitive science society*, 2669–2674.

Goodman, Noah & Michael Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science* 20(11). 818–829. http://dx.doi.org/http://dx.doi.org/10.1016/j.tics.2016.08.005.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics, vol. 3, speech acts*, 41–58. New York: Academic Press.

Jasbi, Masoud, Brandon Waldon & Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology* 10. http://dx.doi.org/https://doi.org/10.3389/fpsyg.2019.00189.

Kao, Justine T, Jean Y Wu, Leon Bergen & Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007. http://dx.doi.org/http://dx.doi.org/10.1073/pnas.1407479111.

Kennedy, Christopher. 1999. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. New York: Garland.

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

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381. http://dx.doi.org/http://dx.doi.org/10.1353/lan.2005.0071.

Klein, Ewan. 1991. Comparatives. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantik: Ein internationales Handuch der zeitgenössischen Forschung* (semantics: an international handbook of contemporary research), chap. 32, 673–691. Berlin: de Gruyter. http://dx.doi.org/http://dx.doi.org/10.1515/9783110126969.8.673.

Lassiter, Daniel & Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory*, vol. 23, 587–610. http://dx.doi.org/http://dx.doi.org/10.3765/salt.v23i0.2658.

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

Lewis, Karen S. 2020a. Metasemantics without semantic intentions. *Inquiry* 0(0). 1–29. http://dx.doi.org/10.1080/0020174X.2020.1847184.

Lewis, Karen S. 2020b. The speaker authority problem for context-sensitivity (or: You can't always mean what you want). *Erkenntnis* 85(6). 1527–1555. http://dx.doi.org/https://doi.org/10.1007/s10670-018-0089-2.

Noveck, Ira A. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78(2). 165–188. http://dx.doi.org/http://dx.doi.org/10.1016/S0010-0277(00)00114-1.

Papafragou, Anna & Julien Musolino. 2003. Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition* 86(3). 253–282. http://dx.doi.org/http://dx.doi.org/10.1016/S0010-0277(02)00179-8.

Pickering, Martin J & Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences* 36(4). 329–347. http://dx.doi.org/http://dx.doi.org/10.1017/S0140525X12001495.

Pinkal, Manfred. 1995. *Logic and lexicon*. Dordrecht: Kluwer. http://dx.doi.org/http://dx.doi.org/10.1007/978-94-015-8445-6.

Qing, Ciyang. 2020. *Semantic underspecification and its contextual resolution in the domain of degrees*: Stanford University dissertation.

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Lucas Champollion & Anna Szabolcsi (eds.), *Proceedings of Semantics and Linguistic Theory 24*, 23–41. http://dx.doi.org/http://dx.doi.org/10.3765/salt.v24i0.2412.

Ronai, Eszter & Ming Xiang. 2020. Pragmatic inferences are QUD-sensitive: an experimental study. *Journal of Linguistics* 57(4). 841–870. http://dx.doi.org/http://dx.doi.org/10.1017/S0022226720000389.

Rotstein, Carmen & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12. 259–288. http://dx.doi.org/http://dx.doi.org/10.1023/B:NALS.0000034517.56898.9a.

Schöller, Anthea & Michael Franke. 2017. Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard* 3(1). http://dx.doi.org/http://dx.doi.org/10.1515/lingvan-2016-0072.

Solt, Stephanie. 2009. Notes on the comparison class. In *International workshop on vagueness in communication*, 189–206. Springer. http://dx.doi.org/http://dx.doi.org/10.1007/978-3-642-18446-8_11.

Solt, Stephanie. 2012. Comparison to arbitrary standards. In *Proceedings of sinn und bedeutung*, vol. 16 2, 557–570.

Syrett, Kristen. 2007. *Learning about the structure of scales: Adverbial modification and the acquisition of the semantics of gradable adjectives*: Northwestern University dissertation.

Tessler, Michael Henry & Noah D Goodman. 2019. The language of generalization. *Psychological review* 126(3). 395. http://dx.doi.org/http://dx.doi.org/10.1037/

rev0000142.

van Tiel, Bob, Michael Franke & Uli Sauerland. 2021. Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences* 118(9). http://dx.doi.org/http://dx.doi.org/10.1073/pnas.2005453118.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. http://dx.doi.org/http://dx.doi.org/10.1093/jos/ffu017.

Toledo, Assaf & Galit Weidman Sassoon. 2011. Absolute vs. relative adjectives - variance within vs. between individuals. In Neil Ashton, Anca Chereches & David Lutz (eds.), *Proceedings of Semantics and Linguistic Theory 21*, 135–154. http://dx.doi.org/http://dx.doi.org/10.3765/salt.v21i0.2587.

Unger, Peter. 1975. *Ignorance*. Oxford, UK: Clarendon Press. http://dx.doi.org/http://dx.doi.org/10.1093/0198244177.001.0001.

Waldon, Brandon & Judith Degen. 2020. Modeling behavior in truth value judgment task experiments. In *Proceedings of the society for computation in linguistics 2020*, 238–247. New York, New York: Association for Computational Linguistics. https://aclanthology.org/2020.scil-1.29.

Ming Xiang
Department of Linguistics
University of Chicago
1115 E. 58th Street
Chicago, IL 60637
mxiang@uchicago.edu

Christopher Kennedy
Department of Linguistics
University of Chicago
1115 E. 58th Street
Chicago, IL 60637
ck@uchicago.edu

Weijie Xu
Department of Language Science
University of California, Irvine
3151 Social Science Plaza A
Irvine, CA 92697-5100
weijie.xu@uci.edu

Timothy Leffel
Seattle, WA
tjleffel@gmail.com

Supplementary material

S1. List of stimuli for Experiment 2 and 3

	Image Type	Image Set	Adjective	Adjective Class
1	artifact	beer	empty	Maximum
2	artifact	beer	full	Maximum
3	artifact	bentnail	straight	Maximum
4	artifact	bentnail	bent	Minimum
5	artifact	boat	big	Relative
6	artifact	boat	small	Relative
7	artifact	book	thick	Relative
8	artifact	book	thin	Relative
9	artifact	bridge	narrow	Relative
10	artifact	bridge	wide	Relative
11	artifact	candle	tall	Relative
12	artifact	candle	short	Relative
13	artifact	chips	closed	Maximum
14	artifact	chips	open	Minimum
15	artifact	curvedbridge	straight	Maximum
16	artifact	curvedbridge	curved	Minimum
17	artifact	fish	plain	Maximum
18	artifact	fish	striped	Minimum
19	artifact	garage	closed	Maximum
20	artifact	garage	open	Minimum
21	artifact	ladybug	plain	Maximum
22	artifact	ladybug	spotted	Minimum
23	artifact	marker	thick	Relative
24	artifact	marker	thin	Relative
25	artifact	noodle	long	Relative
26	artifact	noodle	short	Relative
27	artifact	palm	straight	Maximum
28	artifact	palm	curved	Minimum
29	artifact	pillow	plain	Maximum
30	artifact	pillow	spotted	Minimum
31	artifact	shirt	plain	Maximum
32	artifact	shirt	striped	Minimum
33	artifact	shoe	smooth	Maximum
34	artifact	shoe	bumpy	Minimum
35	artifact	snowman	big	Relative
36	artifact	snowman	small	Relative
37	artifact	sofa	narrow	Relative
38	artifact	sofa	wide	Relative
39	artifact	squash	smooth	Maximum
40	artifact	squash	bumpy	Minimum
41	artifact	stack	tall	Relative
42	artifact	stack	short	Relative
43	artifact	straightrod	straight	Maximum

44	artifact	straightrod	bent	Minimum
45	artifact	table	long	Relative
46	artifact	table	short	Relative
47	artifact	trash	empty	Maximum
48	artifact	trash	full	Maximum
49	shape	bent_bluearrow	straight	Maximum
50	shape	bent_bluearrow	bent	Minimum
51	shape	bent_greenarrow	straight	Maximum
52	shape	bent_greenarrow	bent	Minimum
53	shape	big_redsquare	big	Relative
54	shape	big_redsquare	small	Relative
55	shape	big_yellowcircle	big	Relative
56	shape	big_yellowcircle	small	Relative
57	shape	bumpy_bluesquare	smooth	Maximum
58	shape	bumpy_bluesquare	bumpy	Minimum
59	shape	bumpy_redsquare	smooth	Maximum
60	shape	bumpy_redsquare	bumpy	Minimum
61	shape	curved_blueline	straight	Maximum
62	shape	curved_blueline	curved	Minimum
63	shape	curved-greenline	straight	Maximum
64	shape	curved-greenline	curved	Minimum
65	shape	full_greencube	empty	Maximum
66	shape	full_greencube	full	Maximum
67	shape	full_yellowcube	empty	Maximum
68	shape	full_yellowcube	full	Maximum
69	shape	long_greenarrow	long	Relative
70	shape	long_greenarrow	short	Relative
71	shape	long_greenline	long	Relative
72	shape	long_greenline	short	Relative
73	shape	open_bluecircle	closed	Maximum
74	shape	open_bluecircle	open	Minimum
75	shape	open_redtriangle	closed	Maximum
76	shape	open_redtriangle	open	Minimum
77	shape	spotted_yellowcircle	plain	Maximum
78	shape	spotted_yellowcircle	spotted	Minimum
79	shape	spotted_yellowsquare	plain	Maximum
80	shape	spotted_yellowsquare	spotted	Minimum
81	shape	striped_redcircle	plain	Maximum
82	shape	striped_redcircle	striped	Minimum
83	shape	striped_yellowsquare	plain	Maximum
84	shape	striped_yellowsquare	striped	Minimum
85	shape	tall_greencyclinder	tall	Relative
86	shape	tall_greencyclinder	short	Relative
87	shape	tall_greenspiral	tall	Relative
88	shape	tall_greenspiral	short	Relative
89	shape	thick_bluearrow	thick	Relative
90	shape	thick_bluearrow	thin	Relative
91	shape	thick_redarrow	thick	Relative

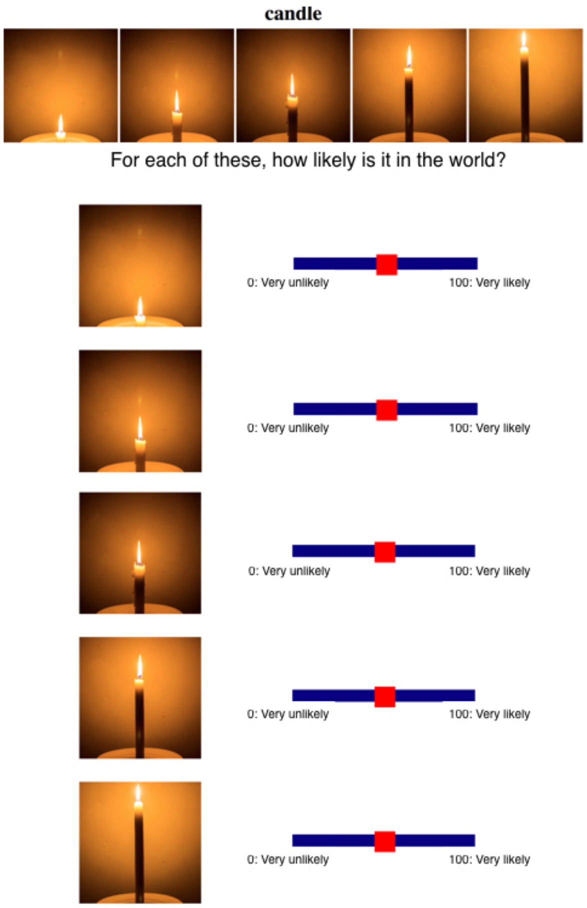
92	shape	thick_redarrow	thin	Relative
93	shape	wide_greenoval	narrow	Relative
94	shape	wide_greenoval	wide	Relative
95	shape	wide_redoval	narrow	Relative
96	shape	wide_redoval	wide	Relative

S2. Using a slider mean rating task to elicit priors

For each of the five objects on the same scale, participants were asked to rate how likely it is in the world[12]. The rating was provided by by moving a bar on a slider that represents a 0 to 100 point range, with 0 corresponding to “very unlikely” and 100 corresponding “very likely.” An example trial is presented in Figure 19. Each participant saw 52 sets of images, including 24 sets of artifact images, 24 sets of shape images, and 4 additional sets of filler trials. The filler trials showed clearly plausible and implausible images, such as normal-looking onions vs. bright pink onions, and they served as attention-check trials.

For each 5-object image set, we obtained the average rating for each scale position and then normalized and transformed the mean rating score for each scale position into a probability value between 0 and 1. After removing participants with poor performance on the attention-check trials, data from 40 participants was included in the analysis. The results are shown in Figure 20. As the figure makes clear, the priors produced by this method are largely flat. Unlike what we found in Experiment 1, these priors did not show any differences between different types of adjective classes, nor did they reproduce any difference between artifacts and shape objects. Using these priors as input to the LG and QF models, with the same parameter values as in the main text, also failed to predict the behavioral (posterior degree and truth value judgments) differences between adjective classes, and between shapes and artifacts, as shown in Figure 21 and 22.

[12] In a different version of the experiment, the instruction in Figure 19 was changed to a more general one “For each of these, how likely is it?", but this did not change the results.



candle

For each of these, how likely is it in the world?

Figure 19: An example trial using the mean slider rating task to elicit degree priors.

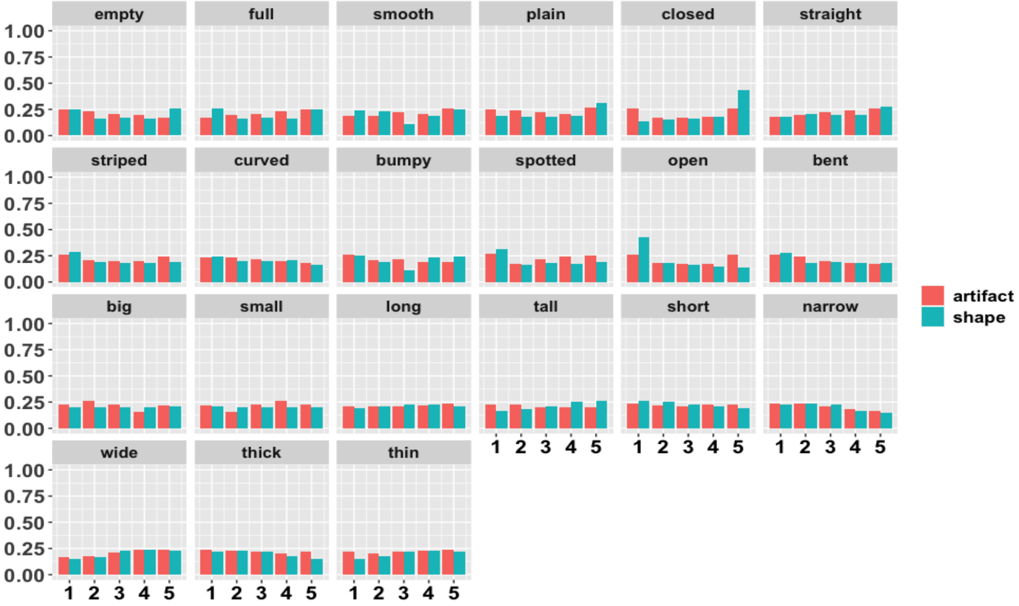


Figure 20: Results of elicitation of degree priors using slider task. Bars represent proportion of selection for each scale position for the image sets used for each adjective in Experiments 2 and 3. **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

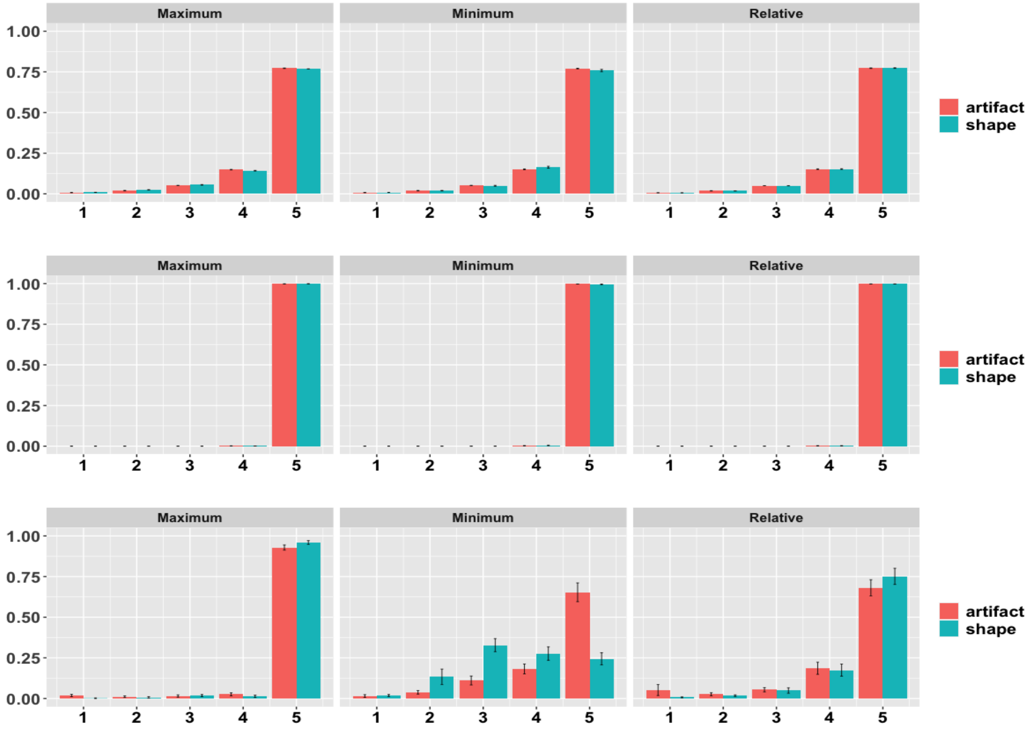


Figure 21: Posterior degree judgments: comparison of LG predictions (top), QF predictions (middle) and empirical judgments (bottom). Model predictions are based on the mean slider rating priors.

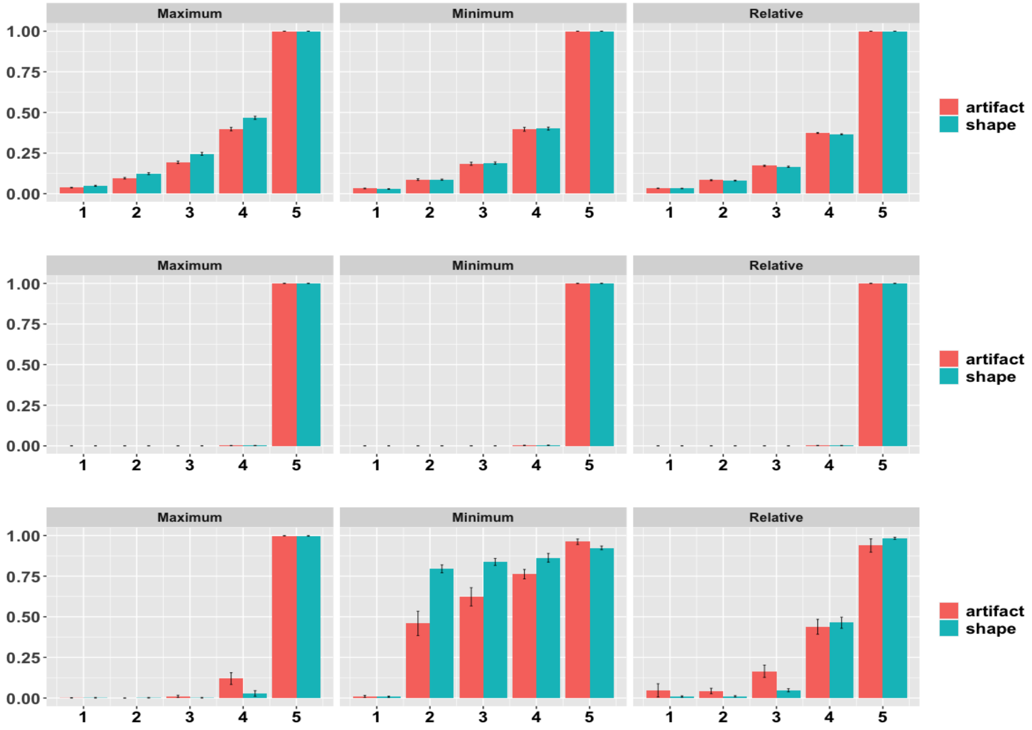


Figure 22: Truth value judgments: comparison of LG predictions (top), QF predictions (middle) and empirical judgments (bottom).Model predictions are based on the mean slider rating priors.

S3. A speaker production-based model for truth value judgments

In this section we extend the basic RSA model in the main text to construct a model of the pragmatic speaker. The current RSA implementation in the main text only included three levels: a literal listener L_0 at the bottom level, a speaker S_1 that reasons about the literal listener, and a pragmatic listener L_1 that reasons about the speaker. But the general RSA framework allows more levels of recursive reasoning. In order to determine whether the empirical truth value judgments of a pragmatic listener could be modeled by the production probabilities of a pragmatic speaker, we need to derive model predictions for the next level pragmatic speaker S_2. If there is a close correspondence between the production choices of a pragmatic speaker and the interpretation of a pragmatic listener, we should observe a correlation between the model predicted production probabilities for S_2 and the empirical human truth value judgments collected in Experiment 2.

Following the basic RSA framework, we construct the pragmatic speaker model as shown in (13): the probability that a pragmatic speaker utters a sentence containing a gradable adjective u_{adj} to communicate that a particular object has a degree d of the relevant adjectival property is based on their reasoning about the trade-off between how a pragmatic listener would interpret the utterance (i.e., the posterior probability that the listener assigns to d upon hearing the utterance) and the cost c of making the utterance:[13]

(13)

$$\begin{aligned} P_S(u_{adj} \mid d) &\propto exp(\lambda(log(P_L(d|u_{adj})) - c)) \\ &\propto P_L(d|u_{adj})^{\lambda} \times exp(-\lambda \times c) \end{aligned}$$

To implement this model, we make the simplifying assumption that a speaker has the option to either utter the target sentence u_{adj} or stay silent $u_{silence}$:

(14)

$$\begin{aligned} P_S(u_{adj} \mid d) &= \frac{P_L(d|u_{adj})^{\lambda} \times exp(-\lambda \times c)}{P_L(d|u_{adj})^{\lambda} \times exp(-\lambda \times c) + P_L(d|u_{silence})^{\lambda} \times exp(-\lambda \times c)} \\ &= \frac{P_L(d|u_{adj})^{\lambda} \times exp(-\lambda \times c)}{P_L(d|u_{adj})^{\lambda} \times exp(-\lambda \times c) + P(d)^{\lambda}} \end{aligned}$$

The term $P_L(d|u_{adj})$ represents a pragmatic listener's posterior belief of the probability of a given degree d upon hearing the adjective u_{adj}. The term $P_L(d|u_{silence})$ represents the listener's belief of the given degree d when the speaker says nothing (i.e. the speaker is silent). The utterance cost is 0 when the speaker is silent, and also the term $P_L(d|u_{silence})$ simply amounts to the listener's prior belief $P(d)$, which can be obtained from the experimental results collected in Experiment 1.

There are already two free parameters in equation 14, λ and cost. Depending on how we derive the listener's posterior degree term $P_L(d|u_{adj})$, additional parameters could be added. Here we consider three options to derive $P_L(d|u_{adj})$. First, we could use the empirically estimated posterior degrees from Experiment 3, in which case no additional parameter is required. Second, we could use the RSA model predicted posteriors $P_L(d|u_{adj})$ (as we have done in section 3.1). In order to derive the model prediction for this term, we need an additional set of parameters λ and cost, yielding a final implementation of (14) with 4 free parameters (2 λs and 2 costs). Alternatively, we could make the reasonable assumption that at least the

[13] The RSA extension of the speaker model in (13) is adapted from Tessler & Goodman (2019), Equation (3) in their paper, with the addition of the utterance cost parameter.

production cost parameter c is shared at different levels of RSA reasoning, and model the ultimate production probability in (14) with 3 free parameters (2 different λs and 1 cost).

We explored all 3 options. For each option, we allowed λ to vary as an integer between 1 and 10, and c as an integer between 0 and 10. We searched through the entire space of possible parameter combinations, and selected the parameter values that yielded the highest R^2 score for the by-item correlation between the model-predicted speaker production probability and the empirical truth value judgments collected in Experiment 2. The highest R^2 scores are as follows:

- 2-PARAMETER MODEL: $R^2 = 0.67$

  $\lambda = 1, c = 0$ the pragmatic speaker level

- 3-PARAMETER MODEL: $R^2 = 0.59$

  $\lambda = 1, c = 0$ the pragmatic listener level

  $\lambda = 5, c = 0$ the pragmatic speaker level

- 4-PARAMETER MODEL: $R^2 = 0.68$

  $\lambda = 2, c = 4$ the pragmatic listener level

  $\lambda = 10, c = 0$ the pragmatic speaker level

The results presented in the Section 3.2.2 were from the 2-parameter model. The averaged results based on adjective classes can be found in Figure 14 in section 3.2.2, and Figure 23 presents results for each individual adjective.

The parameter grid search method described above is relatively coarse-grained. At a reviewer's suggestion, we also conducted an additional search of the parameter space at a finer granularity. For the 2-parameter model, we did a grid search with an increment of 0.1 instead of 1. For the 3-parameter model, we searched for the cost parameter between [0,5] with an increment of 0.2. And for the 4-parameter model, we searched the cost parameter between [0,5] with an increment of 0.5. None of these choices changed the results reported above.

S4. Results for individual adjectives

The following figures provide the empirical results and model predictions for individual adjectives for truth value judgments (Figures 24-26), and posterior degree judgments (Figures 27-29).

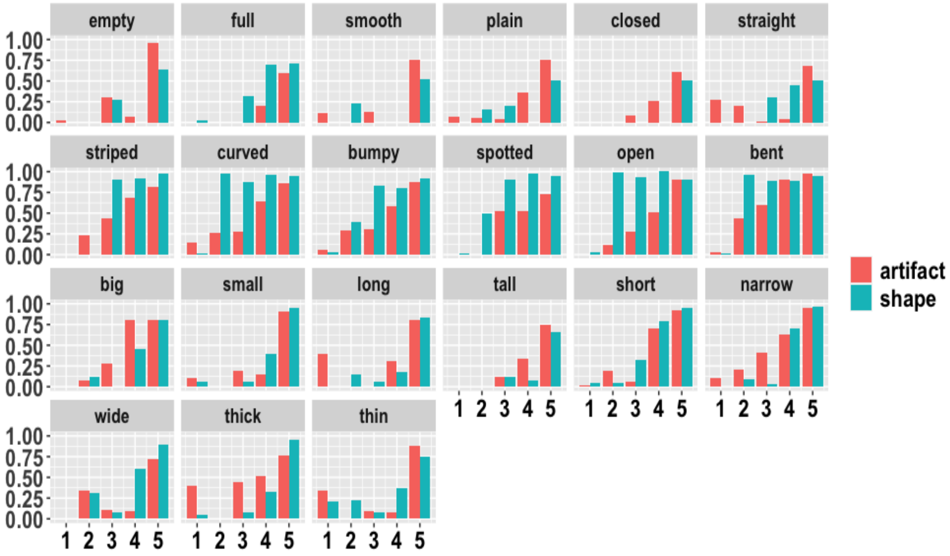


Figure 23: Model predicted by-adjective production probability of a pragmatic speaker. **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

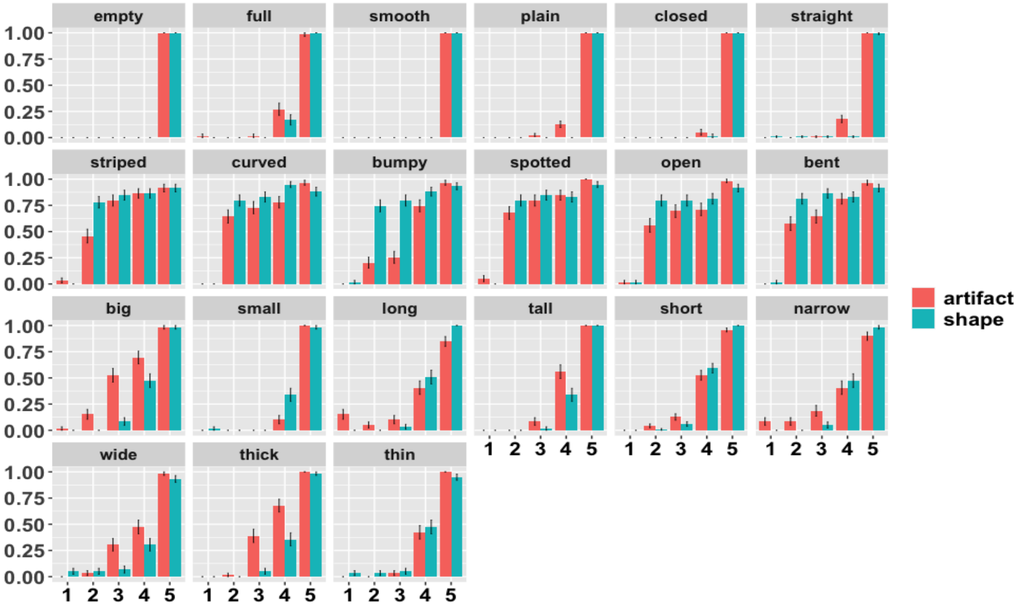


Figure 24: Empirical by-adjective truth value judgments from **Experiment 2** (percentage of positive responses for each scale position). **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

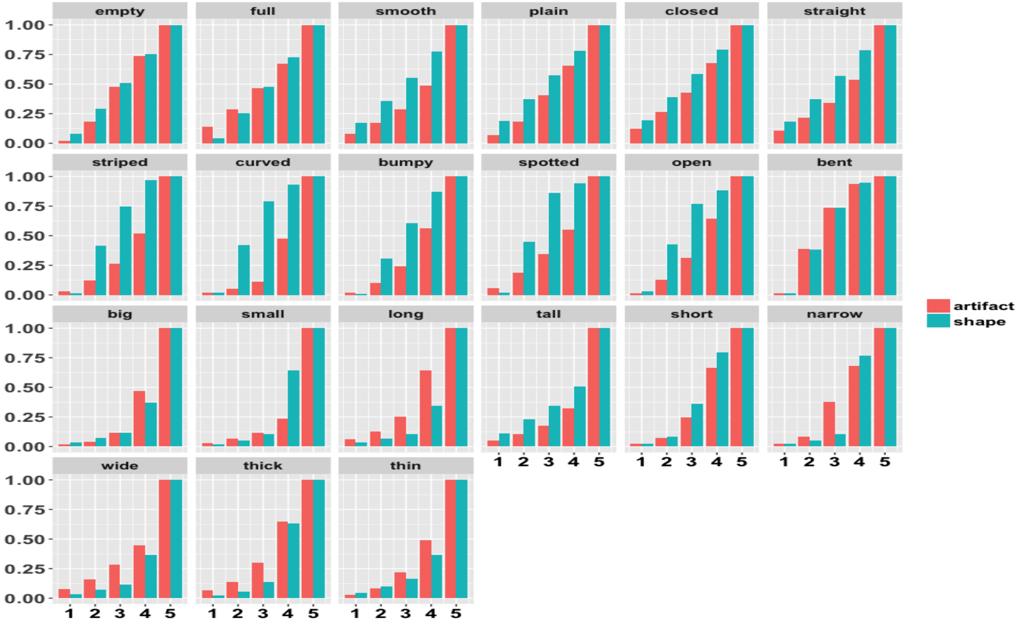


Figure 25: LG model predictions for truth value judgments by adjective.

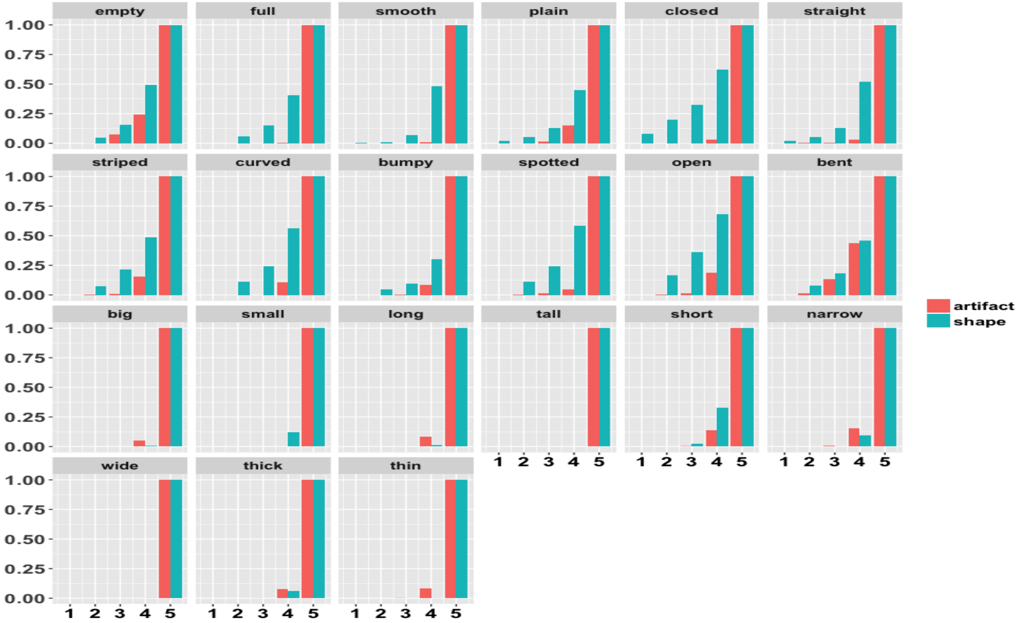


Figure 26: QF model predictions for truth value judgments by adjective.

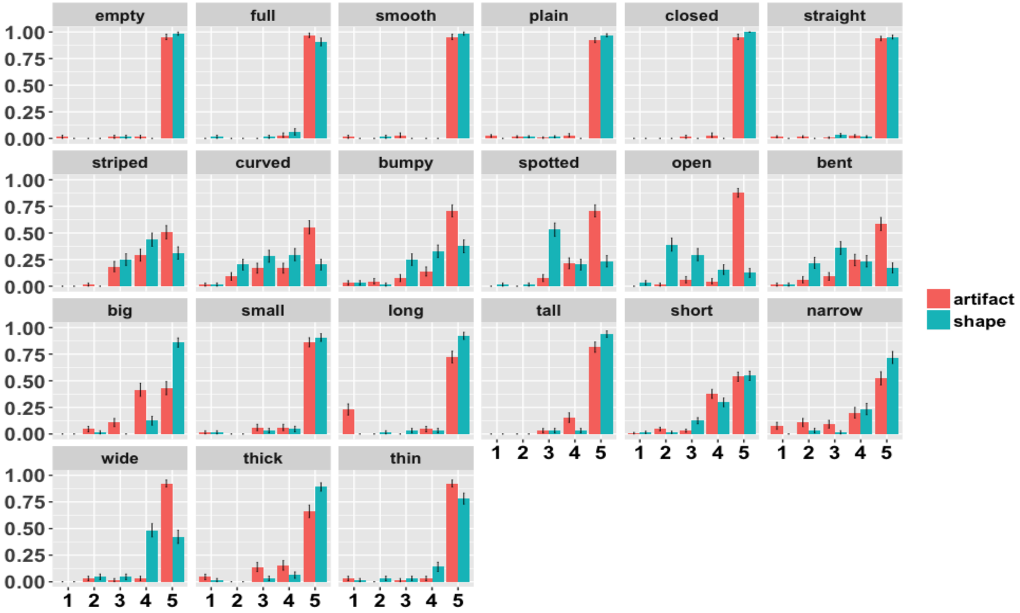


Figure 27: Empirical by-adjective posterior degree judgments from **Experiment 3** (percentage of item selection at each scale position). **Absolute maximum** adjectives are in the top row; **absolute minimum** adjectives are in the second row; and **relative** adjectives are in the third and fourth rows.

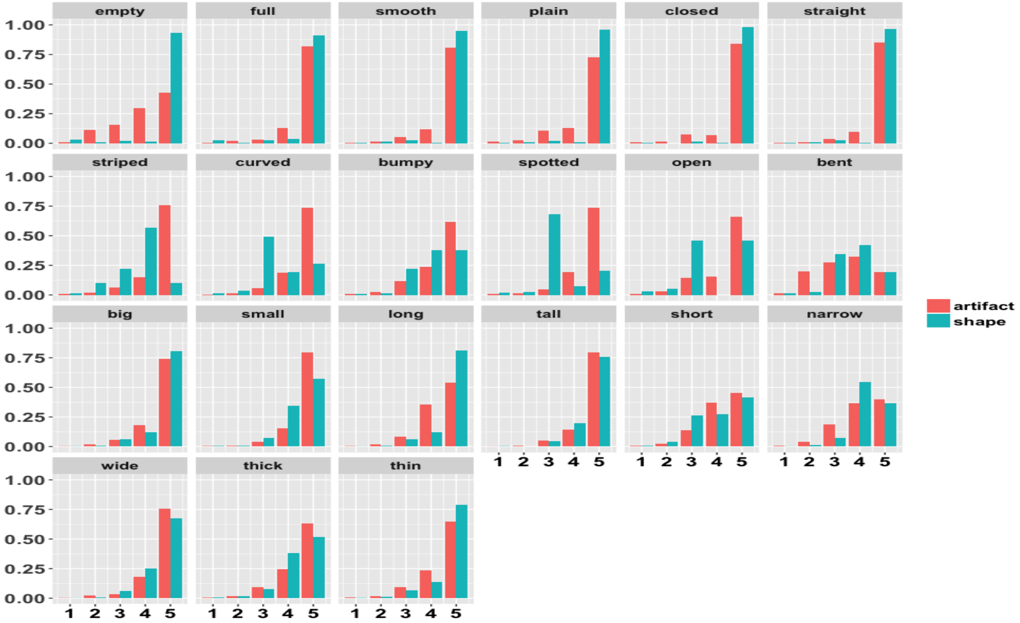


Figure 28: LG model predictions for posterior degrees by adjective.

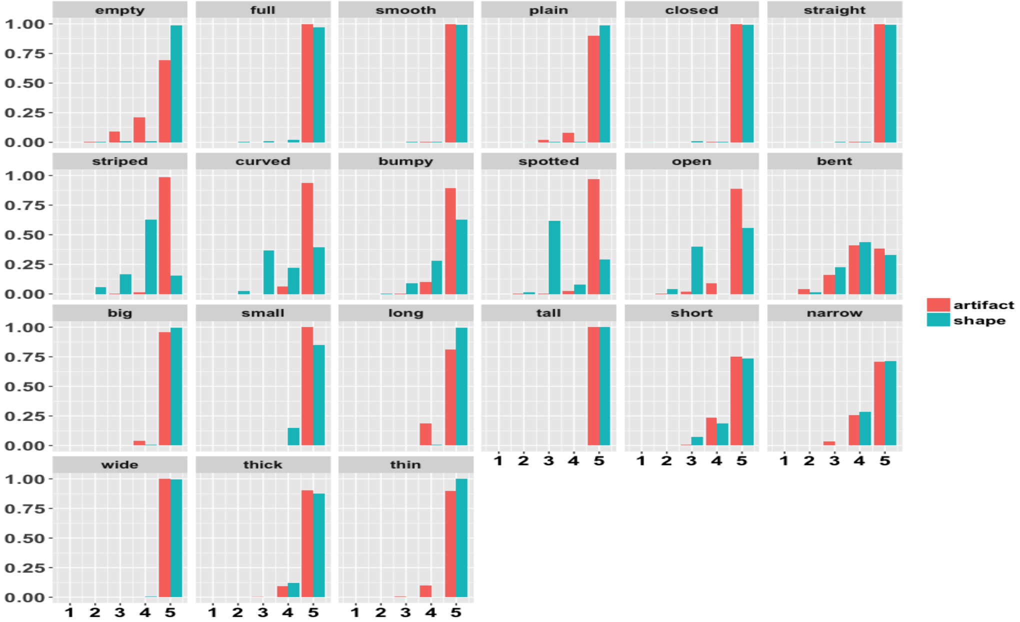


Figure 29: QF model predictions for posterior degrees by adjective.