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

Ming Xiang, Christopher Kennedy,
Weijie Xu and 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 although the Bayesian accounts accurately model human judgments about what is communicated, they do not capture human judgments about truth conditions unless also supplemented with the threshold conventions postulated by semantic theory.

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

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 height; *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

* DRAFT January 20, 2021. A very good day!

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.
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." 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

1 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, with a special model-theoretic "degree" type, or whether it is a contextual parameter relative to which the extension of the adjective is determined, subject to certain consistency constraints. (See e.g. Klein 1991, Kennedy 1999, Burnett 2016 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.

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 e.g. 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 2014, Qing & Franke 2014). 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 his 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 she is 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: we know, for example, that the information communicated by an utterance of a weak scalar term like *some* is often distinct from its truth conditions. The case of gradable adjectives is particularly interesting, however, 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 explicit 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. This can be diagnosed by looking at acceptability with certain types of modifiers (Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007, Syrett 2007). 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]

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.

(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). There are differences in implementation between these accounts, but they share the general feature that closed-scale adjectives default to 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 (2015, 2014) (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 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 a specific implementation of the more general 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 L1, upon hearing an utterance, updates his probabilistic understanding of the world states by reasoning about what a pragmatic speaker S1 could have chosen as her utterances. The pragmatic speaker S1 makes a choice on the utterance by reasoning about a literal listener L0, 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) $P_{L_1}(\ell, \theta \mid u) \propto P_{S_1}(u \mid \ell, \theta) \times P_{L_1}(\ell) \times P_{L_1}(\theta)$

The posterior joint probability of ℓ and θ, for a pragmatic listener, given an utterance u, $P_{L_1}(\ell, \theta \mid u)$, 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, e.g. 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. He updates his 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) $P_{S_1}(u \mid \ell, \theta) \propto exp(\lambda(informativity(u, \ell, \theta) - cost(u)))$

A speaker, in the simplest scenario, could choose to stay silent or make an utterance. The probability of her 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. $informativity = log(P_{L_0}(\ell \mid u, \theta))$
b. $P_{L_0}(\ell \mid u, \theta) = P_{L_0}(\ell \mid [\![u]\!]^{\theta} = 1)$

A speaker therefore evaluates the informativeness of her 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, although an intuitive notion, is actually poorly understood as to how it should be implemented; in the model, *cost* is therefore a free parameter in (9) whose value can be set in different ways (e.g. as a function of the length of the utterance). A second adjustable parameter in (9) is $\lambda \geq 0$, 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 only infer the values of thresholds by considering, for all possible threshold values and all possible literal meanings of the utterance — 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 that the pole in question has a length greater than θ — for various values of θ. The further below the mean of the pole-length distribution that a particular θ is, the more likely it is that an arbitrary pole has at least that much length, and the further above the mean that a particular length 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 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 distribution over the comparison class, and a posterior probability 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 shape of the prior for degrees of pole-straightness is different from the shape of the prior for degrees of pole-length. While the latter is (plausibly) normal, the former is not; instead there is significant probability mass at the upper end of the ordering: we tend to encounter a lot of straight poles. The output of the LG model 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 nail is at or near the maximum. (A minimum standard interpretation of *bent* 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 many 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 the pragmatic reasoning process at the level of a pragmatic listener, thresholds for adjectives are viewed as linguistic conventions learned in a community to achieve optimal communicative efficiency between a speaker and a listener. When a speaker makes a choice between uttering an adjective or saying nothing, she already has probabilistic knowledge $\Pr(\theta)$ about the distribution of θ for a given adjective. The probability of her uttering *"that pole is long"* is determined by comparing the length of the object she is describing and a randomly selected θ from the distribution $\Pr(\theta)$. Once a speaker's production model is specified, a listener in the QF model updates his belief about the length of the same object, conditioned on the utterance, via Bayes' rule in the following way (equation (9) in Qing & Franke 2014):

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

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 value for the correct length in (8).

The critical task for the QF model, then, is to explain and derive the speaker's probabilistic knowledge about 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 his prior belief based on the

utterance of a speaker, he would have the best chance to get the intended degree. Qing & Franke (2014) define an average *expected success rate* of θ, and this further conditions the utility function of θ and the probability distribution of θ ($Pr(\theta)$).[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, "above average," maximal — in virtue of the fact that the objects they are used to order distribute in different ways.

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 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, e.g. beliefs about the lengths or straightnesses of the sorts of poles found at Home Depot. In this sense, QF shares with LG the feature that thresholds are based on prior beliefs about how objects in the world distribute along various dimensions, and not on differences in the lexical semantic representations of different kinds of 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,

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:

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

(ii) $U(\theta) = ES(\theta) - \int_{-\infty}^{\theta} P(\ell).cdh$

(iii) $Pr(\theta) \propto exp(\lambda U(\theta))$

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.

A second difference between the semantic approaches and the Bayesian pragmatic approachs is that the latter but not the former critically make 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 a crucial role in shaping the posterior. In addition, the implementations of the threshold distribution in both models are also heavily dependent on the prior degree distribution. 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 models predictions match the human results for posterior degree estimations but not for truth value judgments; instead, the human truth value judgments correspond closely to the predictions of semantic theories. In section 4 we show that by providing a Bayesian model with θ priors that correspond to truth human truth value judgments — effectively building in the thresholds postulated by the semantic theories — the models (unsurprisingly) make correct predictions about truth values, *and also* retain their correct predictions about posterior degree estimations. We conclude with a discussion of the significance of this result.

2 Empirical estimates from human participants

2.1 Experiment 1: Degree priors

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 a restricted judgment of likelihood as a proxy for prior degree probability: we presented subjects with a set of items that were identical 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 among 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 on the right track; if the judgments are not distinct, we will have reason to think that it is not on the right track.

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. Our inspiration for this distinction comes from a study by Foppolo & Panzeri (2011), 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. In addition to using the shapes/artifacts distinction as a reality check on our methodology for collecting prior degrees, then, we can use it to attempt to reproduce Foppolo and Panzeri's results in a truth value task, extend it to the estimation of posterior degrees, and then ask whether the Bayesian models capture any differences between the two categories that are observed in the human responses, starting from the empirical priors.

Forty-eight sets of images were created, 24 for artifacts and 24 for shapes. Each image set consisted of five items that differed only 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.

The artifacts (24 sets) and shapes (24 sets) were tested in the same experiment. Ninety-seven participants completed the study on IbexFarm; all were self-reported native English speakers recruited from MechanicalTurk. 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, and figure 3 shows the proportion of selection at each scale position, for each

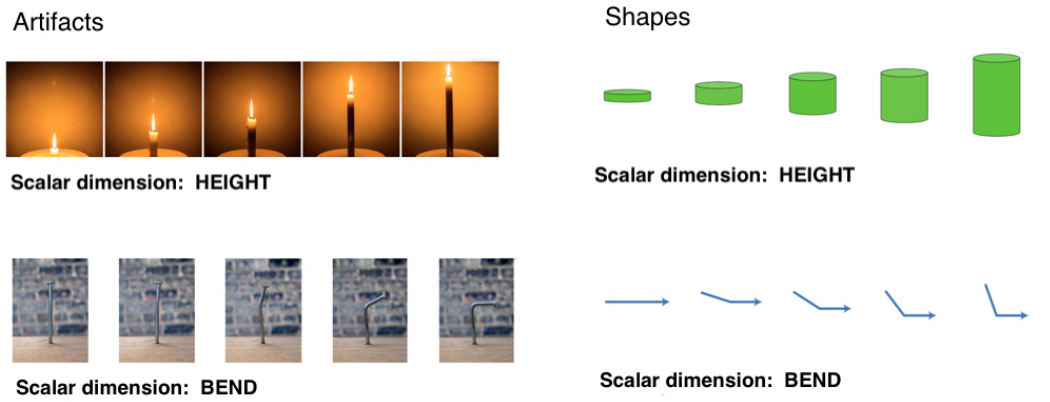

Figure 1: Sample image sets

image set/adjective that was used to elicit judgments about specific adjectives in Experiments 2 and 3.[4]

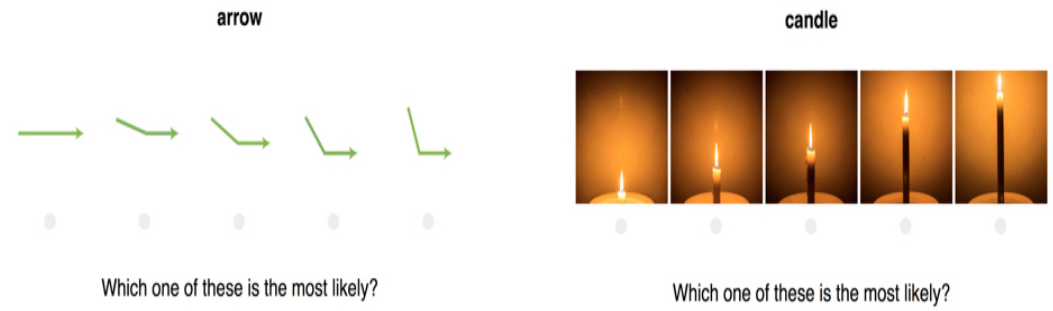

Figure 2: Sample stimuli for Experiment 1: Elicitation of degree priors.

The proportions presented in Figure 3 show how likely it was that a given object at a given scale position was selected. Since the task in this experiment did not involve any linguistic description of either the objects or their properties, and since the objects differed only in the degree to which they manifested the various scalar properties, we believe that it is reasonable to conclude that these measures can be used to stand proxy for participants' beliefs about prior degree distributions for these

4 Since the same image set, e.g. a set of candles of different height, could be used to elicit judgments about different adjectives in Experiment 2 and 3, e.g. *tall* and *short*, prior judgments from the same image set could have contributed to the results of more than one adjective in figure 3.

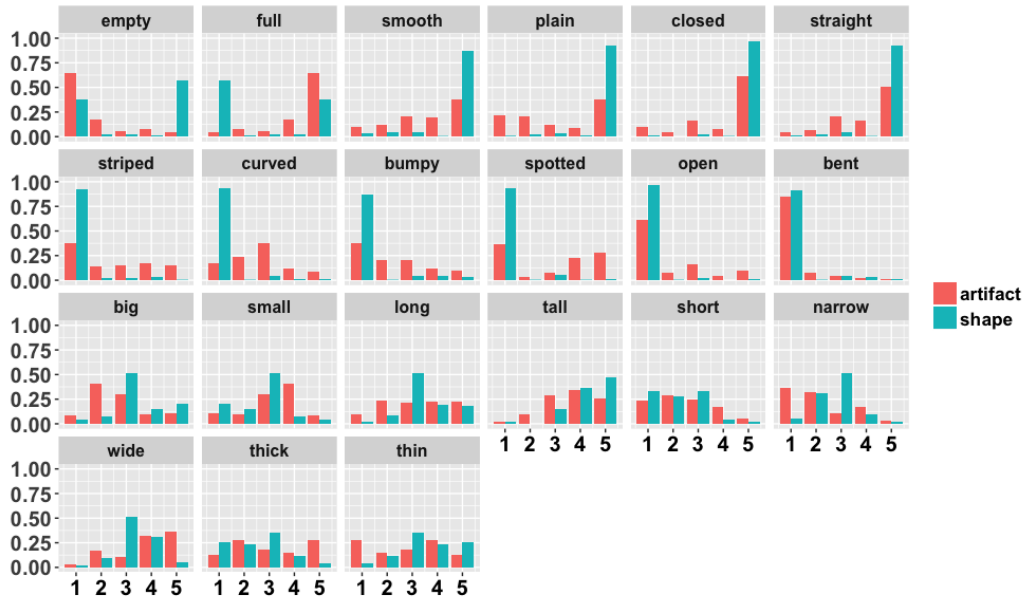

Figure 3: Results of Experiment 1: Elicitation of degree priors. Bars represent proportion of selection for each scale position for the image sets used for individual adjectives in Experiments 2 and 3.

objects. It is also evident from Figure 3 that the pattern of responses for shapes and artifacts were different in the way we expected: artifacts tend to have a more even distribution, while shapes tend to have a more categorical one, in particular for the dimensions corresponding to absolute adjectives. It may turn out that there are better methodologies for collecting priors (a point to which we will return later), but our results suggest that this is a good first step.

2.2 Experiment 2: Truth value judgments

In Experiment 2, the same 48 sets of images that were used for the elicitation of prior degree estimations were used to elicit truth value judgments. As noted above, each image set represented a scalar dimension that could be associated with the members of a par of antonyms. 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*).[5]

The truth value judgment study was conducted on IbexFarm, with all participants recruited from MechanicalTurk. Artifact and shape images were tested separately for two different groups of participants, with 58 self-reported native English speakers in each group. 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. Example stimuli are shown in Figure 4. Since each image set was paired (separately) with each member of an antonym pair, participants in the artifacts group and participants in the shapes group each saw 48 trials in total (24 image sets x 2 adjectives each).

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.[6] Here and elsewhere, the x-axis represents scale positions from low to high degrees. For statistical analysis, we used the R package "brms" to fit Bayesian hierarchical models to the data (Bürkner 2017). 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 model also included by participant and by image_set random intercepts.[7] The

5 *Short* was used twice, both with *long* for the length dimension and with *tall* for the height dimension.

6 Plots for each individual adjective tested, from both the empirical studies and the model predictions, can be found in the Appendix.

7 The model specification is as follows:

- model.formula = response~image_type*adj_class*scale_position + (1+scale|subj) + (1+scale|image_set)

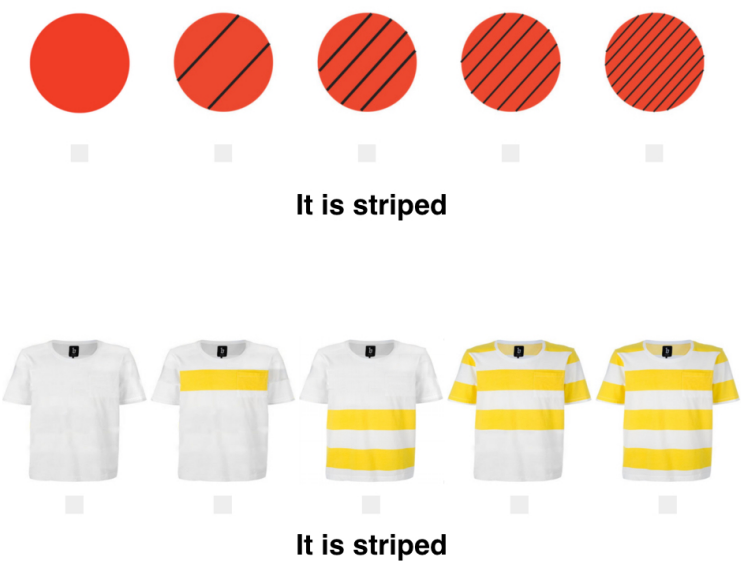

Figure 4: Sample stimuli for Experiment 2: Truth value judgments

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. 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 substantial evidence for an effect.

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

- prior = c(set_prior("normal(0,1.5)", class = "b"), set_prior("normal(0,1)", class = "Intercept"))
- model = brm(formula = model.formula, family = bernoulli(link= "logit"), data = data, prior=prior, iter=6000, control = list(adapt_delta = 0.99, max_treedepth = 12), inits = 0, seed = 500)

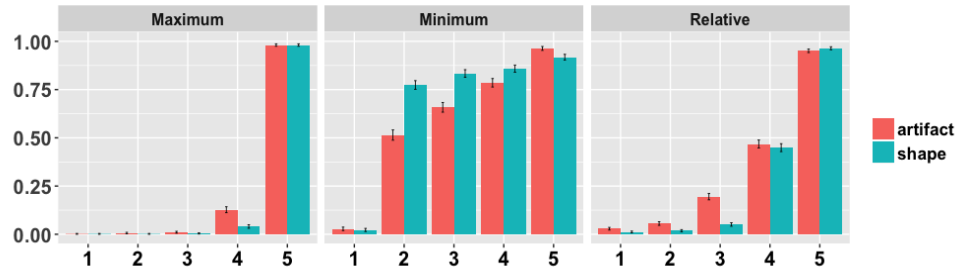

Figure 5: Experiment 2: Percentage of positive responses at each scale position for each adjective class.

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. The responses are further modulated by scale positions. With every unit increase in scale position, there is a greater increase of positive responses for maximum adjectives compared to the relative ones, and a lesser increase of positive responses for minimum adjectives compared to the relative ones. 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).

Effects	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	-2.76	0.24	-3.23	-2.28
Maximum Adj	-6.02	0.35	-6.71	-5.35
Minimum Adj	4.01	0.22	3.59	4.44
Maximum x Scale	3.32	0.25	2.84	3.82
Minimum x Scale	-1.25	0.16	-1.58	-0.94
Image type	0.47	0.22	0.03	0.90
ImageType x Maximum	0.85	0.34	0.17	1.53
ImageType x Minimum	-1.10	0.21	-1.51	-0.68

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

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. To better understand the effect of image type on each adjective class, we carried out analysis for each adjective class separately. For each adjective class, we looked

at the effect of image type and scale positions[8]. Image type was treatment coded with artifact as the baseline level, and scale position was centered before entering into the model. The results from these models are presented in Table 2. For the maximum adjectives, there is no clear 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 prior to scale position 5. For the minimum adjectives, there is strong evidence that participants gave more positive responses for the shape objects than for the artifact objects. There is no clear evidence for an interaction between image type and scale position, although Figure 5 suggests that with the increase of scale positions, the difference between artifacts and shape objects decreases. For the relative adjectives, there is evidence for both an effect of image type and an interaction between image type and scale position. There are fewer positive responses for the shape objects than for the artifact objects, but that difference decreases with the increase of the scale position.

Maximum adjectives	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	-1.04	1.14	-3.25	1.22
Image type	-0.76	1.23	-3.17	1.69
Scale position	2.47	0.80	0.86	3.99
Image type x Scale position	0.44	0.79	-1.12	1.99
Minimum adjectives				
Intercept	1.68	0.59	0.54	2.84
Image type	2.76	0.82	1.16	4.39
Scale position	3.13	0.33	2.49	3.80
Image type x Scale position	0.80	0.46	-0.10	1.69
Relative adjectives				
Intercept	-2.65	0.41	-3.47	-1.86
Image type	-1.28	0.55	-2.35	-0.18
Scale position	2.92	0.32	2.30	3.55
Image type x Scale position	1.09	0.44	0.23	1.95

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

8 For each adjective class, model.formula = response~image_type*scale_position + (1+scale|subj) + (1+scale|image_set); other parameters were set as above.

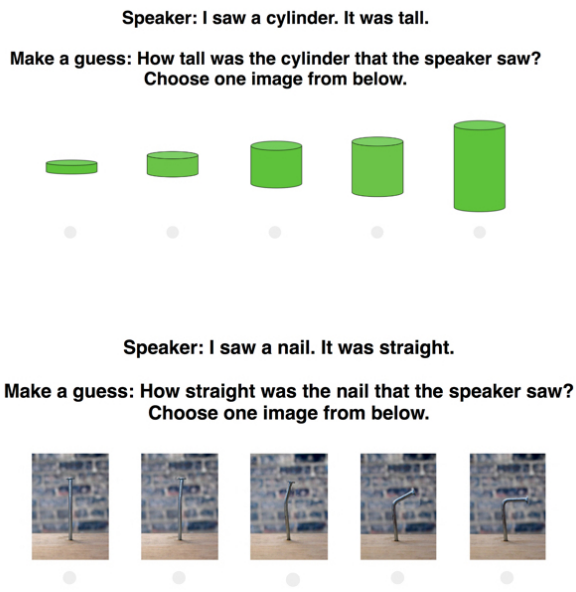

Figure 6: Sample stimuli for Experiment3: Estimating degree posteriors

2.3 Experiment 3: Posterior degrees

In our final experiment, we 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. We used the same set of images as in Experiments 1 and 2, but 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 same five objects from the image sets used in Experiments 1 and 2. Example stimuli are shown in Figure 6.

The experiment was conducted on IbexFarm. As in Experiment 2, shape images and artifact images were tested in two different sessions. Sixty-seven participants were recruited for the shape session, and a different group of sixty-eight participants for the artifact session. The average percentages of choices for each scale point are presented in Figure 7.

Upon visual inspection, the averaged results in Figure 7 show the following qualitative patterns: for maximum adjectives, participants consistently chose the

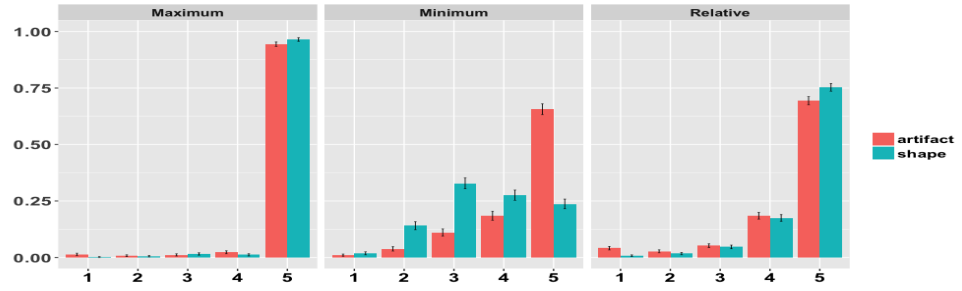

Figure 7: Experiment 3: Percentage of item selection at each scale position for each adjective class.

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. 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. In Table 3) we present results when all data are considered together. and in Table 4 we present results from analyses performed for each adjective class.

Effects	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	-3.65	0.23	-4.11	-3.20
Maximum Adj	-3.10	0.33	-3.74	-2.45
Minimum Adj	1.34	0.24	0.87	1.82
Maximum x Scale	2.21	0.23	1.75	2.67
Minimum x Scale	-1.21	0.18	-1.59	-0.86
Image type	0.18	0.23	-0.28	0.63
ImageType x Maximum	0.50	0.32	-0.13	1.13
ImageType x Minimum	-1.15	0.24	-1.62	-0.68

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, based on the empirical degree priors

Maximum adjectives	**Estimate**	**SE**	**Lower 95% CrI**	**Higher 95% CrI**
Intercept	-1.42	1.08	-3.51	0.72
Image type	-0.25	1.16	-2.54	2.01
Scale position	0.87	0.87	-0.91	2.53
Image type x Scale position	0.31	0.84	-1.33	1.95
Minimum adjectives				
Intercept	-2.66	0.21	-3.07	-2.26
Image type	0.99	0.26	0.47	1.52
Scale position	1.58	0.19	1.21	1.95
Image type x Scale position	-1.09	0.25	-1.58	-0.59
Relative adjectives				
Intercept	-3.20	0.35	-3.88	-2.48
Image type	-0.83	0.49	-1.78	0.15
Scale position	1.93	0.29	1.35	2.48
Image type x Scale position	0.65	0.40	-0.14	1.42

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

we collected in Experiment 1.[9] The model predictions for posterior degrees can be directly compared to the empirical estimates of posterior degrees collected in Experiment 3, but as noted in section 1.3, the Bayesian models do not actually make predictions about truth value judgments. They do, however, make predictions about threshold posteriors, which can be transformed into predictions about truth value judgments as described below and compared to the empirical results from Experiment 2.

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 adjectival scale we tested in the experiments, and we likewise derived the model predictions for posterior threshold distribution for each adjective scale, also distributed over the five points on the 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.

The basic procedures for running these two models were as follows. The inputs to each model were the empirically estimated priors from Experiment 1. The LG model also requires priors for the threshold θ, which we set to be a uniform distribution, as in Lassiter & Goodman 2014. As described in Section 1.2.2 (equation (9)), both models also make use of two free parameters, λ and *cost*. The results presented below were based on λ=3 and *cost* = 2, which are comparable to the values used by Lassiter & Goodman (2014) and Qing & Franke (2014); variations within reasonable range did not qualitatively change the model outcomes. (We will say more about the procedure of selecting the values for these two parameters in section 4). When running the two models, we first generated predictions for each image set that was used in the empirical studies, and then these results were averaged to derive predictions for each individual adjective (when one adjective was used to describe different image sets) and for each adjective class.

3.1 Posterior degrees

Figure 8 shows the LG and QF model predictions for posterior degrees, alongside the empirical posteriors collected in Experiment 3, for each adjective class.

9 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.

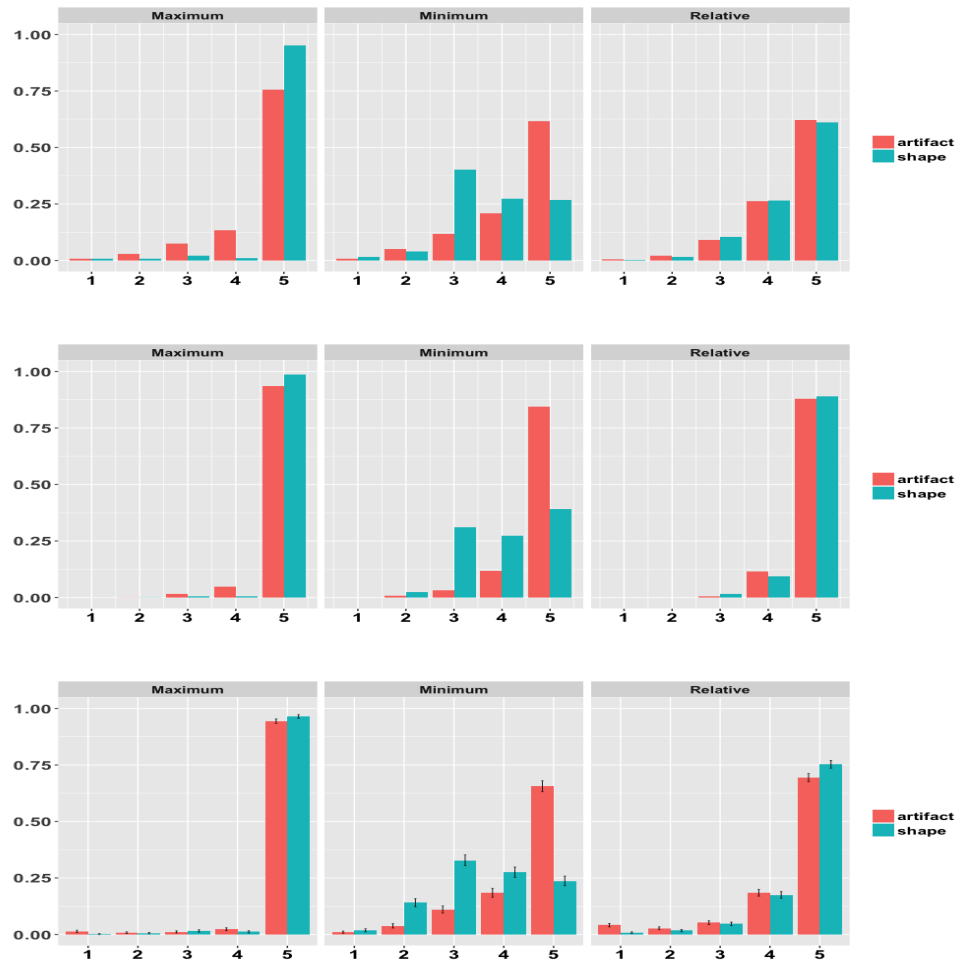

Figure 8: Comparison of LG predictions (top), QF predictions (middle) and Experiment 3 empirical estimates (bottom) of posterior degrees.

Visual inspection of these figures suggests that both models make good predictions, with the overall patterns of the model predictions quite similar to the actual patterns in the empirical data. Qualitatively speaking, 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, scale point 4 also has a non-trivial amount of

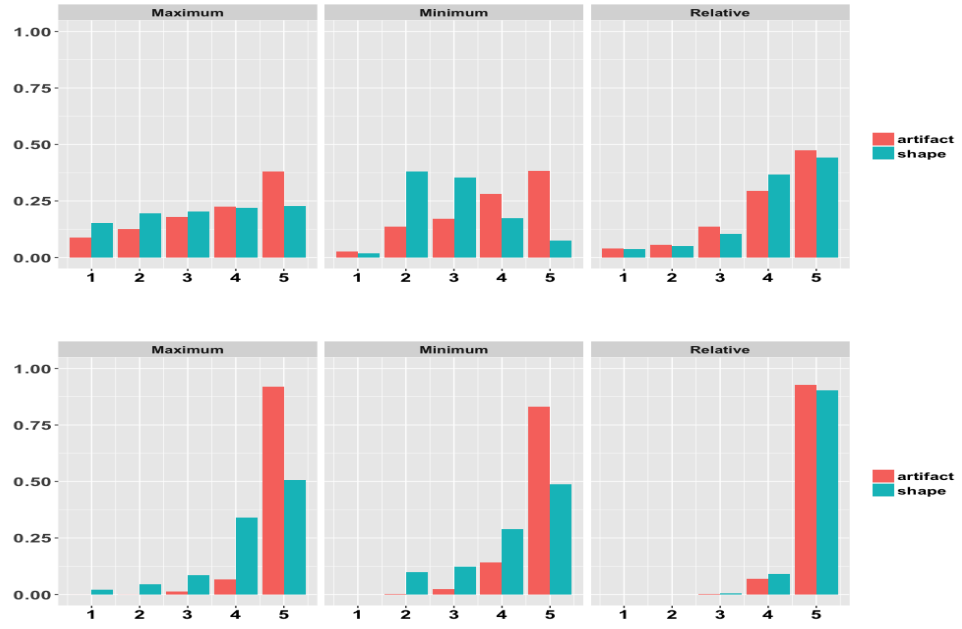

Figure 9: LG (top) and QF (bottom) predictions for thresholds, derived based on the empirical priors estimated in Experiment 1.

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.

3.2 Truth value judgments

As noted above, the Bayesian models do not make direct predictions about truth value judgments. However, they do make predictions about posterior threshold distributions, which are shown in Figure 9 for each adjective class. Based on the model predictions for threshold distribution, we can derive predictions about truth value judgments by adopting the following hypothesis about the link between beliefs about threshold probability and truth value judgments. 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 of the object is a function of her belief about the likelihood that the object's height is at least as great as the threshold: the more likely she believes the threshold to be no greater than the height of the object, the more likely she is to judge the utterance as true. Given this relationship, we can derive model predictions for the probability that an utterance would be judged true of an

object at scale position i by summing up the the predicted posterior probabilities of thresholds at each scale position $j \leq i$:

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

Figure 10 shows the LG and QF model predictions for truth value judgments for each adjective class that are derived in this way, compared to 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 the endpoint-oriented truth conditions of these adjectives. 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 exactly 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. 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, it appears that both the LG and the QF models fail to accurately predict human truth value judgments for absolute adjectives. The LG model does well with minimum adjectives, but it treats the maximum adjectives as if they were minimum. 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 as through they were maximum ones.

4 General discussion

In Section 3.1 we saw that, using empirically derived prior degree distributions as input, the Bayesian pragmatic models of communication with gradable adjectives generate predictions about posterior degree distributions that are qualitatively similar to human estimations. However, their predictions about truth value judgments do not match human responses. Since the model predictions for truth value judgments are based on a mapping from threshold posteriors to truth value judgments, the relatively poor model performance arises entirely from the model predictions for threshold distributions.

One potential explanation for the models' poor performance on threshold predictions and truth value judgments is that it reflects a problem with our methodology

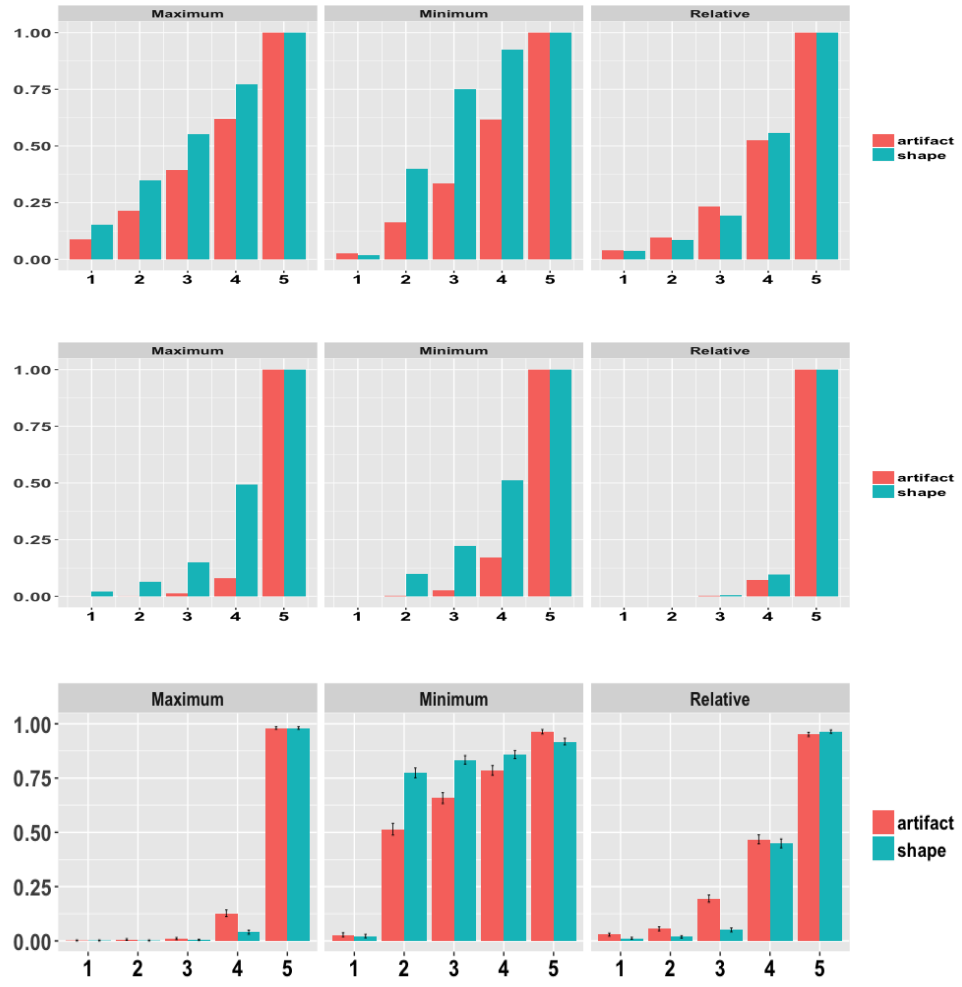

Figure 10: Comparison of LG predictions (top), QF predictions (middle) and Experiment 2 empirical judgments (bottom) of truth values.

for deriving empirical degree priors. If this were the case, however, then we should also see poor model performance on degree posteriors, but this was not the case.

A second potential explanation for the models' poor performance is that we were not using optimal values for the free parameters λ and *cost*. To test this possibility, we formed 100 different (λ, *cost*) settings by fixing each variable to an integer value between 1 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 setting, the model predictions were correlated with the experimental data from Experiment 2 and 3 to derive a R^2 score, which we used to evaluate the model fit. Across the R^2 scores obtained from the 100 (λ, *cost*) settings, the best models in general have a better model fit for the posterior degree judgments than for the truth value judgments. Prior to considering any adjective class distinctions, for the LG model, the R^2 values range between 0.32 and 0.86 for the posterior degree judgments, and between 0.38 and 0.67 for the truth value judgments. For the QF model, the R^2 values are between 0.75 and 0.88 for the posterior degrees, and between 0.53 and 0.64 for the truth value judgments. There is larger variability for the R^2 scores associated with the LG model, suggesting the LG model predictions are more sensitive to changes in the choices of the free parameter values, whereas the QF model predictions are relatively more stable. 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 this set of parameter values, we further calculated the model fits for each adjective class separately. The results are shown in Table 5.

	LG (R^2)		QF (R^2)	
	TVJ	POSTERIOR	TVJ	POSTERIOR
Overall	0.65	0.84	0.6	0.88
Maximum	0.63	0.94	0.85	0.98
Minimum	0.7	0.67	0.39	0.71
Relative	0.91	0.81	0.81	0.87

Table 5: With $\lambda = 3$ and *cost* = 2, correlations between the LG and QF model predictions and the experimental results, for each experiment and each adjective class

Overall, there is good model fit, for both models, for the relative adjective class. The QF model also does well on the maximum adjective class. Both models fall somewhat short for the minimum adjective class. By and large though, both models make better predictions for the posterior judgments than for the TVJ judgments, in line with the descriptive conclusion we drew in section 3.2. In short: Bayesian pragmatic models are good at predicting posterior degree distributions that match

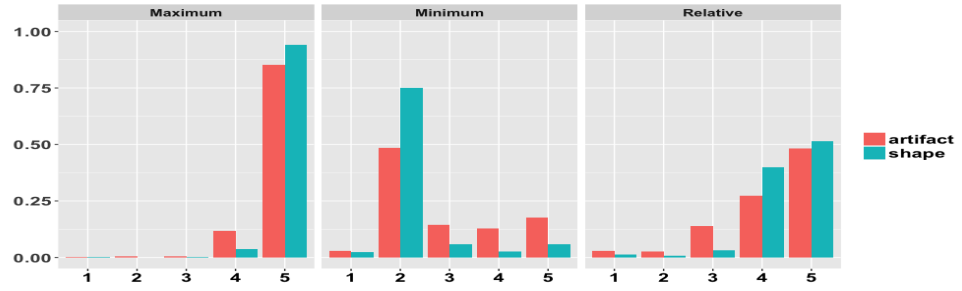

Figure 11: Threshold distribution based on empirical truth value judgments in Experiment 2.

human judgments based on empirical priors, but not so good at predicting threshold distributions and the corresponding truth value judgments.

This leads to the following question: what kind of threshold distributions *would* 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 Figure 5 and the bottom panel of Figure 10) and our hypothesized link between thresholds and truth value judgments (the equation in (12)) to "reverse engineer" a threshold distribution from the empirical responses. The result of this computation is shown in Figure 11.

In Figure 11, maximum adjectives have thresholds that are largely aligned with the maximum degree on the scale. Thresholds for artifacts show a small probability to fall on scale point 4, but thresholds for shapes have a stricter maximum-endpoint orientation. For minimum adjectives, scale point 2 is the most likely threshold, with shapes having a stronger preference for a strict minimum-endpoint orientation than artifacts. Relative adjectives show a more gradual pattern of threshold distribution, and thresholds for shapes again show a more categorical distribution than those for artifacts. The "reverse engineered" threshold distributions in Figure 11 are clearly distinct from the distributions predicted by the Bayesian models, which suggests that these models — or at least their current implementations — are not adequate for deriving gradable adjective thresholds and corresponding truth value judgments. This is perhaps not surprising, since, as we noted in Section 1.3, the Bayesian models are designed to answer question (P): what is the information communicated by an utterance involving a positive form gradable adjective? On the other hand, the threshold distribution in Figure 11 is very close to what we would expect from theories designed to answer question (S) — what are the truth conditions of utterances involving a positive form gradable adjective? — i.e., theories that

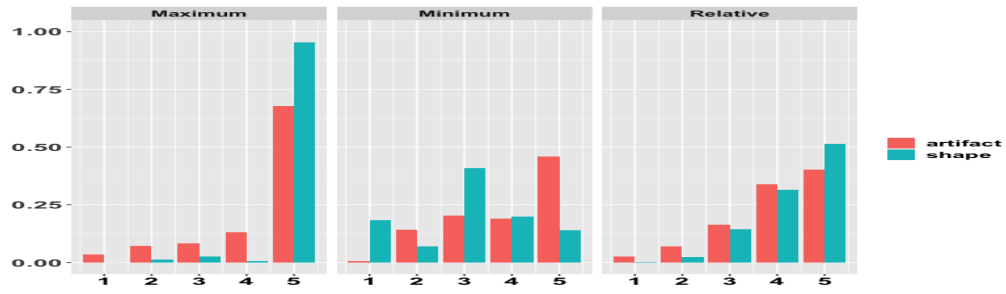

Figure 12: Posterior degree predictions of QF model incorporating "semantic" thresholds.

postulate semantic conventions for associating adjectives that use closed scales with endpoint-oriented thresholds.

Given that the semantic theories provide a good answer to (S) while the Bayesian pragmatic theories provide good answers to (P), but not the other way around, we can ask whether a model that incorporates both the semantic and Bayesian pragmatic theories can do a better job of capturing the full pattern of human responses. To answer this question, we took our implementation of the QF model, and modified it by replacing the original calculation of the threshold distribution (described in section 1.2.2) with the distribution in Figure 11, and used this together with the empirical priors we collected in Experiment 1 to derive a new set of predictions for posterior degrees. The result of this computation is shown in Figure 12.

If we compare these results with the posterior degree predictions made by the original QF and LG models and the empirical estimates of posterior degrees collected in Experiment 3, shown respectively in the top, middle and bottom panels of Figure 8, we can see that the modified model with "semantic" thresholds is on a par with the original models in capturing the empirical posterior degree results. In short, introducing semantic thresholds does not weaken the models' capacity to successfully predict what is communicated, even when that information is strictly weaker (for absolute maximum adjectives) or stronger (for absolute minimum adjectives) than what the truth judgments based on those thresholds would lead us to expect.

Our interpretation of this result is not that absolute adjectives require a "two stage" analysis, with a semantic account of thresholds and a distinct pragmatic model for communicated content: after all, the threshold distribution in Figure 11, while distinct from the threshold distributions derived from empirical priors, nevertheless shows an influence of empirical priors in the more categorical distribution of shape thresholds compared to artifact thresholds. Instead, we take this result to indicate that a single pragmatic model can successfully account both for communicated

content (posterior degree estimations) and truth judgments (threshold estimations), provided that it is supplemented by reasoning based on lexical semantic content, in particular the scale structures of particular adjectives. We know from studies of human language development that such information is both available in the input and attended to by children in the acquisition process (Syrett 2007), so it makes complete sense that it should be integrated into a pragmatic model of communication with gradable adjectives.

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 very good job at capturing human judgments about the degree information that gradable adjectives communicate, something that traditional semantic analyses cannot explain, but they do not adequately capture human judgments about truth conditions. Although it is surely the case that prior beliefs about how the objects in an adjective's domain distribute along a scalar continuum play a role in determining thresholds — this is the only option for relative adjectives, and it is the most plausible explanation for differences in truth value judgments about artifacts vs. shapes for absolute adjectives — our results indicate that this information alone is insufficient to to accurately compute thresholds for absolute adjectives. When the priors for absolute adjectives are restricted to values around scalar endpoints, however, effectively combining empirical priors with lexical semantic information and introducing a semantic convention for associating close-scale adjectives with endpoint-oriented thresholds, Bayesian pragmatic models are able to simultaneously capture *both* human judgments about truth conditions and human judgments about what is communicated.

References

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.

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.

Goodman, Noah & Michael Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science* 20(11). 818–829.

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.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381.

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.

Lassiter, Daniel & Noah Goodman. 2014. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of semantics and linguistic theory*, vol. 24, 587–610.

Lassiter, Daniel & Noah D Goodman. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* 1–36. http://dx.doi.org/10.1007/s11229-015-0786-1.

Pinkal, Manfred. 1995. *Logic and lexicon*. Dordrecht: Kluwer.

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.

Rotstein, Carmen & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12. 259–288.

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

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.

Unger, Peter. 1975. *Ignorance*. Oxford, UK: Clarendon Press.

Appendix

The following tables provide the empirical results and model predictions for individual adjectives, for both the truth value judgments and the posterior degree judgments.

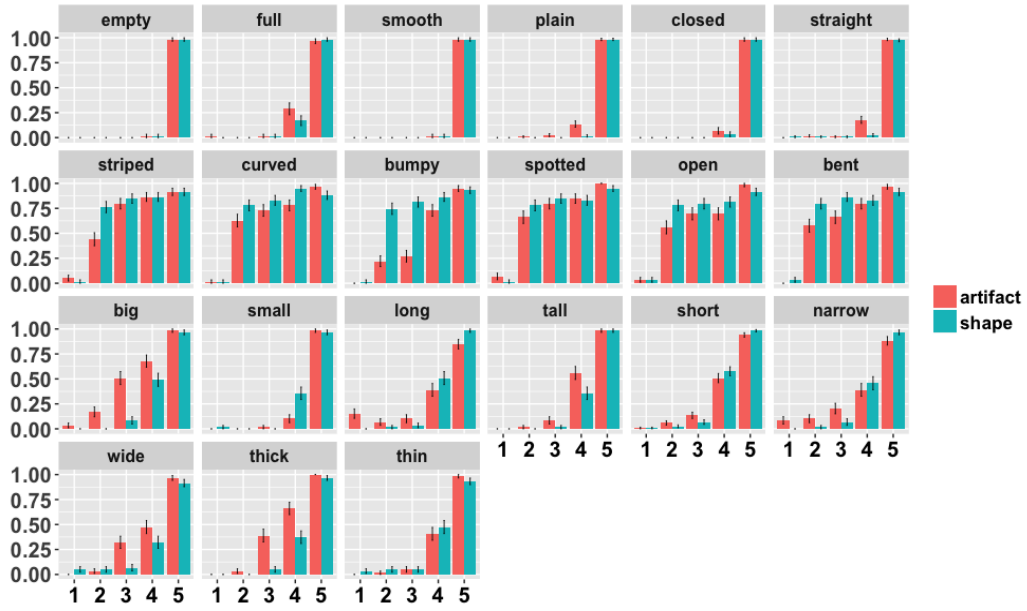

Figure 13: Empirical truth value judgments (percentage of positive responses for each scale position) by adjective.

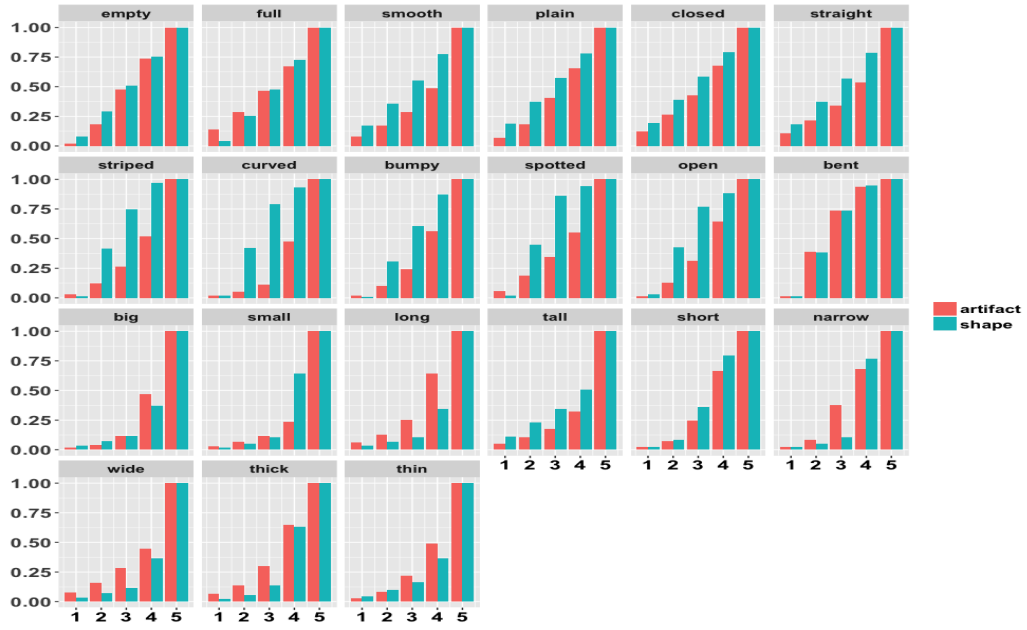

Figure 14: LG model predictions for truth value judgments by adjective.

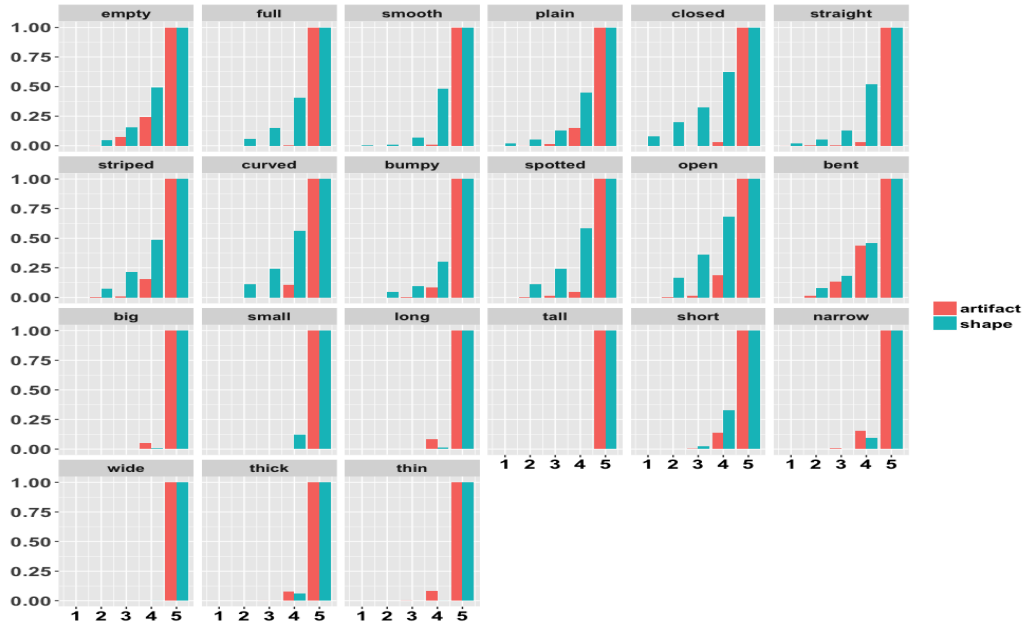

Figure 15: QF model predictions for truth value judgments by adjective.

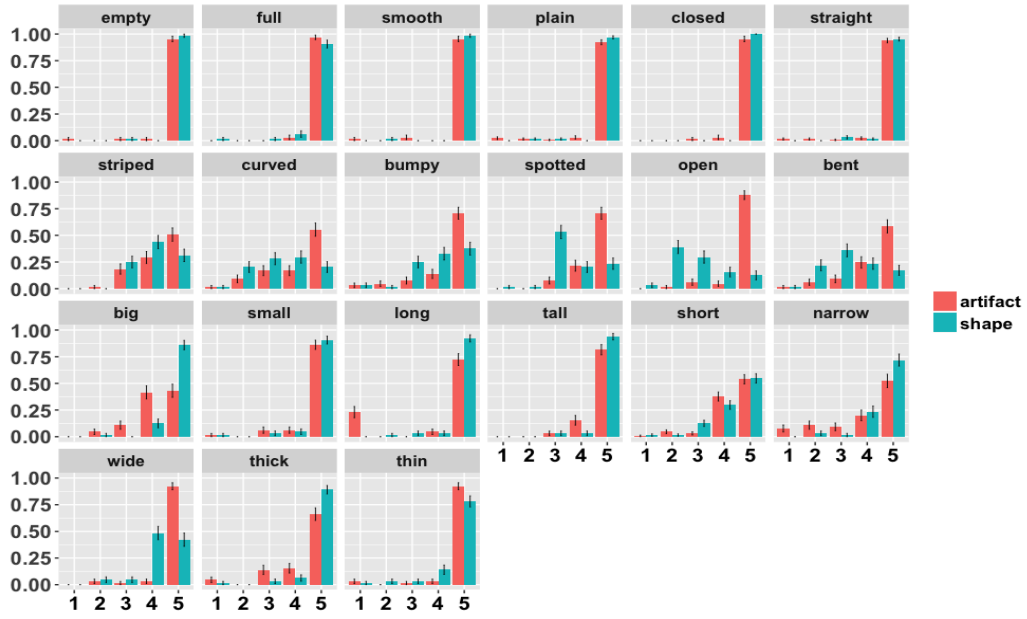

Figure 16: Empirical posterior degree judgments (percentage of item selection at each scale position) by adjective.

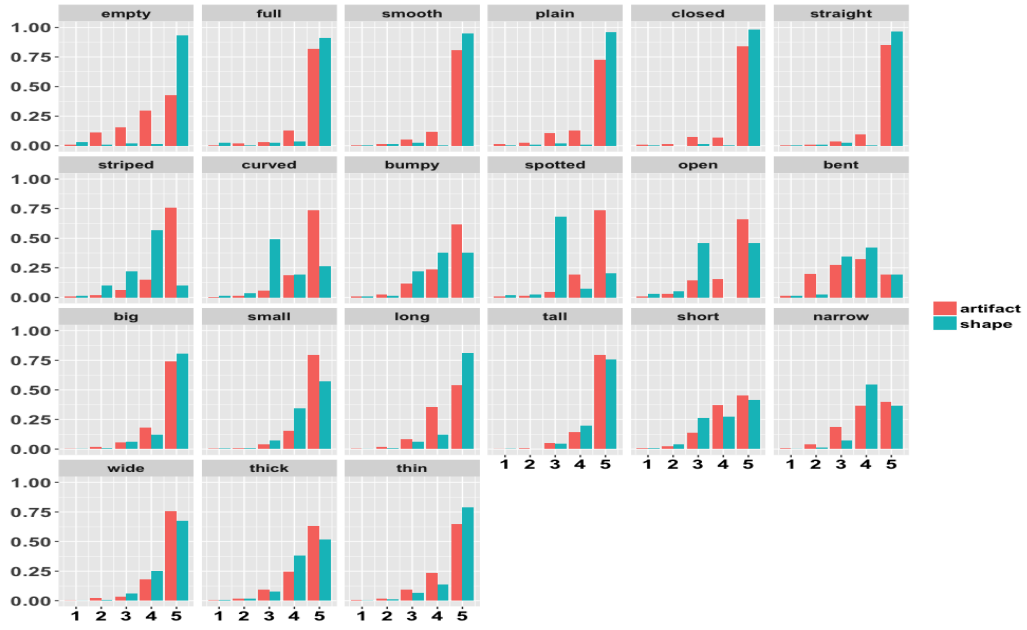

Figure 17: LG model predictions for posterior degrees by adjective.

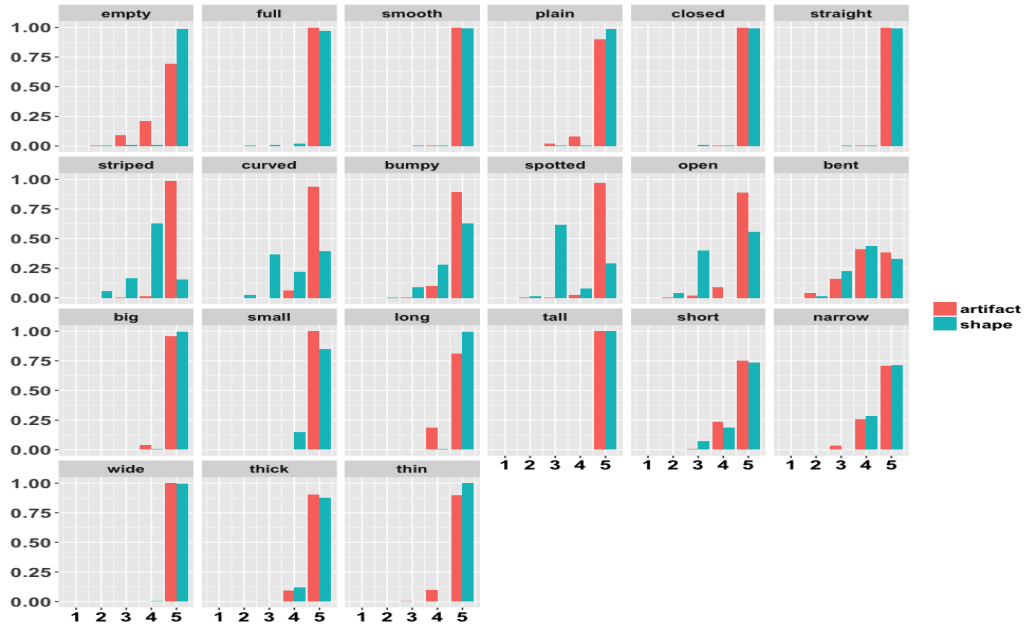

Figure 18: QF model predictions for posterior degrees by adjective.