Interpreting Gradable Adjectives in Context: Domain Distribution vs. Scalar Representation*

Timothy Leffel, Ming Xiang and
Christopher Kennedy
University of Chicago

Abstract Gradable adjectives denote properties that are relativized to context dependent 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 the typical case, thresholds are uncertain, a feature that distinguishes them from other kinds of contextual parameters. Threshold uncertainty has played a prominent role in discussions of vagueness in philosophy of language, but there has been relativelly little attention to this feature of adjective meaning in linguistic semantics and pragmatics, and correspondingly little attention to a fundamental question about communication with gradable adjectives: what are the principles that govern decisions about threshold values in context, and so determine the communicative content of gradable adjectives in a speech situation? In this paper, we consider two recent answers to this question, which differ in the role that formal properties of adjective meaning play in determining threshold values. On one view, decisions about thresholds are based entirely on probabilisitic prior knowledge of how the objects in a gradable adjective's domain distribute relative to the scalar property it expresses (Lassiter & Goodman 2014, 2015); according to the second view, a formal property of adjective meaning — scale structure — is primarily responsible for threshold determination (Kennedy 2007, Potts 2008, Qing & Franke 2014a,b). We identify different predictions of the two approaches and present the results of an experimental study that supports the representational approach over the distributional one.

Keywords: absolute gradable adjectives, Bayesean pragmatics, context dependence, relative gradable adjectives, scale structure, semantic uncertainty

1 Threshold (un)certainty in gradable adjective interpretation

Gradable adjectives are adjectives that support orderings of the objects in their domains relative to some scalar concept. The gradable adjectives *long* and *short* order relative to height; *heavy* and *light* order relative to weight, and so on. Non-gradable

* DRAFT June 29, 2017.

adjectives like *digital* and *next*, on the other hand, are not associated with a scalar concept, at least not grammatically. Grammatical gradability is reflected in both meaning and syntactic distribution: gradable adjectives can appear in comparative constructions (*longer*, *less heavy*) while non-gradable adjectives cannot (**more digital*, **less next*), for example.

There are different formal characterizations of gradability in the literature, and of the difference between gradable and non-gradable adjectives, some of which posit a type-theoretic distinction between the two kinds of expressions and some of which do not. One feature that all analyses agree on, however, 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.
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

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.

the threshold is "determined by context." And indeed, it is evident that property expressed by a gradable adjective in the positive form is context dependent in a way that is consistent with the idea the the threshold may be different in different contexts. (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 for what counts as long, such that 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 other context dependent expressions are analyzed, however, which presents a challenge to the naive view that the semantic value of the degree argument of a positive form gradable adjective is "determined by context." 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 puchasing 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 deny (2a), 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 such as *long, heavy* and *big*, which have inherently context dependent and uncertain thresholds, are often referred to as RELATIVE gradable adjectives. As it turns out, however, not all gradable adjectives have inherently uncertain thresholds. Alongside relative adjectives stands a class of ABSOLUTE gradable adjectives, which often involve threshold uncertainty of some sort, but also have uses in which there is certainty about the threshold (Unger 1975, Pinkal 1995, Rotstein & Winter 2004, Kennedy & McNally 2005, Kennedy 2007, Toledo & Sassoon 2011, Lassiter & Goodman 2014, Qing & Franke 2014a). 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. The pole is straight.
b. The theater is empty.
c. The 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. The pole is bent.
b. The door is open.
c. The shirt is striped.

Note that the claim is not that absolute adjectives are not context dependent in some way, nor that they always involve certainty about thresholds; rather it is that they show a more limited range of context dependence than relative adjectives, and that they have uses in which there is certainty about the threshold. 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.

The possibility of maximal or minimal thresholds constitutes a basic semantic distinction between gradable adjectives: some adjectives encode scalar concepts that are based on closed scales, while others encode scalar concepts that are based on open scales. This distinction 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]

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

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

That open scale adjectives have relative interpretations is unsurprising: the scalar concepts these adjectives encode are consistent with a potentially infinite range of threshold values (there is no inherent limit to the set of lengths, weights, etc.), and provide no basis for privileging one threshold over another: in Williamson's (1992) terms, open scales such as height, weight and so forth lack "natural transitions" that might otherwise provide salient points for coordination on thresholds. That closed scale adjectives have absolute interpretations, however, is a fact that requires

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.

explanation, because having a minimal or maximal value is a necessary condition for an absolute interpretation but not a sufficient one. Two kinds of explanations for this correlation have been proposed in the literature, which differ in the role that the scalar properties of gradable adjectives play in determining threshold values, and more generally in the extent to which formal properties of linguistic expressions play a role in the determination of the communicative content of utterances that introduce semantic uncertainty. In the first approach, decisions about threshold value are based entirely on probabilisitic prior knowledge about the pattern of distribution of objects in a gradable adjective's domain over its scale (Lassiter & Goodman 2014, 2015). Scale structure can influence such patterns, but plays no direct role in threshold determination. In the second approach, the endpoints of closed scales provide coordination points for default maximum or minimum thresholds (Kennedy & McNally 2005, Kennedy 2007, Potts 2008, Toledo & Sassoon 2011, Qing & Franke 2014a,b). On this view, a representational feature of gradable adjective meaning — scale structure — plays the central role, with prior knowledge of distributional patterns or other pragmatic factors modulating default interpretations.

In this paper, we present an experimental study designed to adjudicate between the distributional and representational approaches to threshold determination. In section 2, we describe the key assumptions of the two approaches in more detail, showing how they capture core properties of relative and absolute adjectives and also where they make diverging predictions. Specifically, we show that they differ in their predictions about the interpretation of absolute adjectives in contexts involving impoverished prior knowledge of domain distribution. In section 3 we present an experimental study designed to test these predictions, the results of which support the representional approach over the distributional one. Section 4 concludes with discussion of the larger implications of the results for our understanding of the principles that determine the communicative content of expressions that introduce semantic uncertainty.

2 Capturing the relative/absolute distinction

2.1 Domain distribution

Lassiter & Goodman (2014, 2015) (LG) develop a model of gradable adjective interpretation that starts from what is arguably the null hypothesis about the basic semantics of relative and absolute gradable adjectives: since both classes of adjectives can combine with expressions that compositionally manipulate thresholds, such as the comparison and sufficiency constructions in (1), and since both can have context dependent interpretations in the positive form, their core meanings are identical. Specifically, they both denote threshold-sensitive properties of the sort described

above, and when θ is not determined compositionally, as is the case in positive form predications, its value is indeterminate. Crucially, in this approach, nothing about the grammar or the lexical content of absolute vs. relative adjectives specifies or prioritizes a particular way of resolving this indeterminacy, so any intepretational differences between the two classes of adjectives must be attributed to considerations outside of the semantics proper. These considerations, according to Lassiter and Goodman, involve a general communicative strategy for resolving uncertainty about variables. In the LG model, these principles are stated in terms of Bayesian theory of cooperative communication, in which a "pragmatic listener" computes the a probability distribution for the value of an uncertain variable in an utterance as a function of a relevant set of prior values and the assumption that speakers choose partciular utterances (over relevant alternatives) with (at least) a goal of optimizing informativity.

In the case of gradable adjective interpretation, the LG model model uses prior knowledge of the distribution of the degrees to which the objects in a gradable adjective's domain — or a contextually determined subset of such objects, called a COMPARISON CLASS — possess the scalar concept associated with the adjective to derive a probability distribution on thresholds, and a corresponding posterior probability about the scale position of an object that is described as having the property expressed by the adjective. As illustration, consider an utterance of (8a), whose interpretation we will characterize using the representation in (8b).[3]

(8) a. This nail is long.
b. $\mu_{\mathsf{length}}(\mathsf{n}) \succeq \theta$

Suppose that the comparison class consists of the sorts of nails that are encountered in typical household carpentry contexts, the lengths of which are approximately normally distributed. The further below the mean that a particular length is, the more likely it is that an arbitrary nail 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 nail has that length. Given the speaker's preference for informativity, a low value for θ (e.g. one that makes *long* true of 75% of 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 *long* true of only 1% of the comparison class) will also be assigned a low probability, because the resulting meaning would be too strong. Bypassing a number of details, the output of the LG model in a simple case like this

3 Here and elsewhere, $\mu_{\mathsf{s}}(\mathsf{x})$ represents the value of x on scale s (x's length, emptiness, bend, etc.) and θ is a variable over values on the same scale. These representations are designed to transparently illustrate the threshold-dependent truth conditions of sentneces containing positive form gradable adjectives, not to endorse a particular model-theoretic approach to adjective denotations.

is a posterior probability distribution over 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. The end result is that (8a) is heard to communicate something roughly equivalent to "the length of this nail is significantly greater than the average length of nails in the comparison class," which is an accurate paraphrase of its truth conditions, and crucially provides informative informational content while at the same time retaining a certain degree of uncertainty about the actual lenth of the nail.

In the case of a predication involving an absolute adjective like *straight*, the semantics delivers exactly the same kind of interpretation: (9a) expresses the a relation between the straightness of the nail and a threshold variable, whose value must be determined in context.

(9) a. This nail is straight.
b. $\mu_{\mathsf{straight}}(\mathsf{n}) \succeq \theta$

The reason that (9a) is generally heard to communicate something stronger than "the straightness of this nail is significantly greater than the average straightness of nails in the comparison class" (which would count a lot of bent nails as straight) is because the prior distribution of degrees of straightness for objects in the comparison class is different from the prior distribution of degrees of length in a crucial way: there is significant mass on the upper end. 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. (The 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.

Because Lassiter and Goodman's model assumes that both relative and absolute adjectives express relations to context dependent thresholds, it clearly can account for the fact that both kinds of adjectives have context dependent uses. The differences between the two classes of adjectives are the result of a general decision procedure for assigning posterior probabilities to threshold valuations, and emerge from differences in prior knowledge about how objects in an adjective's domain are distributed relative to the scalar concept the adjective describes. Relative interpreations arise in the case of priors with a relatively normal distribution and little probability mass at the

extremes; absolute interpretations arise in the case of priors that involve a clustering of probability mass at the upper or lower end of the distribution. The relative/absolute distinction is therefore:

> "...not a binary distinction, but a matter of degree: interpretations may adhere more or less strongly to an endpoint, depending on the form of the prior. Nor is it a property of adjectives per se: rather it is a property of interpretations which emerges from the interaction of an adjective's lexically determined scale structure and interpreters' background knowledge about the scalar property which the adjective describes." (Lassiter & Goodman 2014: 601-2)

But it is important to point out that scale structure plays only an indirect role. Closed scales allow for the possibility of a clustering of probability mass around the minimal or maximal endoints of the scale while open scales do not, so the kind of domain distribution that will promote an absolute intepretation is available only in the former case. But closed scales are in principle compatible with normal or flat distributions, and in such cases the result should be a relative or relative-like interpretation. More specifically, the less that prior knowledge about distribution supports a clustering of probability mass at scalar endpoints, the greater uncertainty there will be about how close to maximal the threshold should be, leading to a more relaxed and relative-like interpretation.

2.2 Scalar representation

In contrast to the distributional approach to threshold determination, the representational approach provides scale structure with a direct role in threshold determination, such that closed scales give rise to certainty about (maximal or minimal) thresholds as a default, while open scales give rise to "informative uncertainty" about thresholds of the sort that we saw with the distributional approach. There are a number of different implementations of this idea, but their unifying feature is the hypothesis that the representational structure of closed scales gives rise to a convention for default maximal or minimal thresholds because such thresholds optimize the communicative utility of adjectives that are compatible with them (i.e., the closed-scale ones), given the assumption that the truth conditions of gradable adjectives in general involve relations of the sort shown in (8b) and (9b) (see e.g. Kennedy 2007, Potts 2008, Toledo & Sassoon 2011; see also Burnett 2016 for a similar way of cashing out this idea in a semantics for gradable adjectives that does not involve scalar representations).

This idea is worked out in most detail by Qing & Franke (2014a), who provide a Baysean model for adjective interpretation that is similar in many respects to

the LG model (and derives comparable results for relative adjectives), but differs in incorporating the idea of an optimal linguistic convention from evolutionary approaches to meaning, which prioritizes maximal or minimal thresholds whenever there is any probablity mass at scalar endpoints at all — i.e., whenever an adjective has a closed scale — independent of the prior distribution of degrees for the objects in a comparison class. In effect, the Qing-Franke model uses the scale as the prior rather than the degree distibution for the comparison class when doing so is informative, and derives threshold certainty as a default. However, it is flexible enough to incorporate the influence of world knowledge or factors about particular contexts of utterance to influence decisions about thresholds. If the context justifies using a particular degree distribution rather than the scale structure as the prior — e.g. the straightness of pool cues in dive bars or the number of people who typically attend a theater performance — this can lead to a relaxation of the default threshold and a more uncertain, relative-like interpretation.

2.3 Predictions

The distributional and representational accounts of the relative/absolute distinction make similar predictions about the way that rich priors can give rise to threshold uncertainty and context dependent interpretations of absolute adjectives; where they differ is in their predictions about the interpretive effects of impoverished prior knowledge about the way that the objects in the comparison class distribute relative to the scalar property associated with a particular adjective.

The distributional theory predicts that impoverished knowledge about priors should lead to greater uncertainty about thresholds for absolute adjectives and a weakening of the interpretive distinction between absolute and relative adjectives. In particular, in the simple case of a uniform prior, which is a reasonable assumption only for closed-scale adjectives and represents the absence of substantive prior distributional knowledge, the Lassiter and Goodman model predicts an increase in threshold uncertainty. The model still delivers a preference for thresholds near the scalar endpoints, but with "more slack than with end-peaked priors" (see Lassiter & Goodman 2014: p. 601, figure 3), which corresponds to a more relative-like interpretation.[4] The representational theory, in contrast, predicts that impoverished prior knowledge about domain distribution should reduce any motivation to move

4 Following McNally (2011), Lassiter and Goodman point to uses of *full* to describe wine glasses that are conventionally filled only partway as an example of such a case. While we agree that there are relative-like uses of absolute adjectives — see our examples with *empty* and *straight* in section 1 — we do not think this case is one of them. Consider the following statement in a context in which the convention is to fill wine glasses to the 2/3 mark.

(i) My glass is only half full.

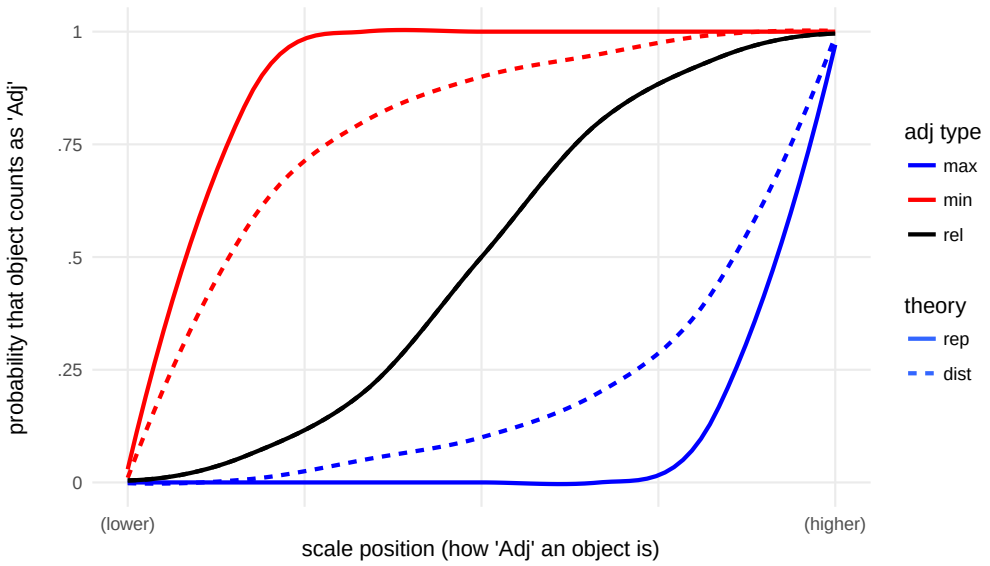

Figure 1: Predictions of distributional and representational theories of the relative/absolute distinction, in the case when knowledge about domain distribution is impoverished.

away from the representation-based coordination on a maximal or minimal threshold, leading to greater certainty about the threshold and a sharpening of the interpretive distinction between absolute and relative adjectives.

These divergent predictions about interpretation make corresponding predictions about behavior, specifically about the interaction between an object's location along a scalar dimension, and the probability that a speaker will judge an absolute adjective to be true of it. These predictions are illustrated in Figure 1: the horizontal axis represents the position of an object on a scale; the vertical axis represents the probability that an object is judged to have the property expressed by the adjective. For any scale point, the greater the certainty that it exceeds or fails to exceed the threshold — which is a function of certainty about the threshold itself — the greater the certainty with which the object will be judged to have, or not have, the property expressed by the adjective. The black line in Fig. 1 represents an idealized relative interpretation pattern, in which judgments are (more or less) directly proportional to scale position. The dashed red and blue lines represent the predictions of the distributional approach for minimum and maximum absolute adjectives, respectively,

(i) is ambiguous: it can either mean that the wine in the glass reaches the 1/2 mark, or it can mean that the wine in the glass reaches the 1/3 mark. The availability of the latter interpretation shows that what counts as a scalar maximum for *full* need not be determined by the physical properties of the container, but can also be determined by conventions for filling a container.

in the case of impoverished prior knowledge of domain distribution. The solid red and blue lines represent the predictions of the representational theory in the same situation. As the graph makes clear, the distributional theory predicts a more relative-like pattern of judgment for absolute adjectives in context involving impoverished priors, while the representational theory predicts an even more categorically absolute pattern.

3 Experiment

To test the predictions of the distributional and representational approaches to threshold determination and the relative/absolute distinction, we conducted an experimental study designed to compare interpretations of relative, absolute minimum, and absolute maximum adjectives using objects for which we can directly manipulate rich vs. poor prior knowledge of domain distribution. Our experimental design owes its inspiration to a study reported by Foppolo & Panzeri (2011), who found a difference in the interpretation of absolute adjectives depending on whether they were predicated of objects with well-defined and familiar comparison classes, like bananas and roads, or of objects without such comparison classes, like random lengths of rope. Foppolo and Panzeri found that the former types of objects systematically allowed relative-like interpretations of absolute adjectives, while the latter types of objects resisted such interpretations. In our study, we further sharpened the distinction between object type by comparing judgments about pictures of familiar, everyday objects, about which we can assume rich prior knowledge of domain distribution, to judgments about about artificially-constructed images of geometric shapes like cubes, pentagrams, and arrows, which we assume to involve impovershed priors.

3.1 Methods and Materials

3.1.1 Design

We collected judgments about the application of gradable adjectives to objects with varying locations on the adjectives' scalar dimensions. Research subjects were shown pictures of objects and asked to judge whether the object had a property expressed by a gradable adjective, using a design adapted from Kim et al. (2014). Stimuli differed along four dimensions: the semantic classification of the adjective as relative, maximum standard absolute, or minimum standard absolute (**adjective type**); the position of the displayed object in an ordering determined by the adjective (**scale position**); whether the object was an abstract shape or a familiar artifact (**object type**); and whether the object was presented in isolation or alongside a family of similar objects with differing scale positions (**presentation type**). Presentation type

and object type were manipulated between subjects, yielding four groups of subjects corresponding to each combination of the levels of presentation type and object type. This design allowed us to investigate the effects of rich/impoverished prior knowledge of objects (via object type), as well as the immediate presence/absence of a salient comparison class (via presentation type), on judgments about different types of gradable adjectives (adjective type) evaluated at different locations along scalar dimensions (scale position).

3.1.2 Participants

Research participants were recruited either from undergraduate linguistics classes at The University of Chicago (in which case they participated for research awareness credit), or from flyers and email solicitations (in which case they were compensated $10). 57 subjects completed the experiment in one of the two shapes conditions (isolated or grouped presentation). Three were excluded due to a non-English native language, leaving 54 sets of shapes responses. 60 subjects completed the experiment in one of the two artifacts conditions. Three were also omitted due to non-native English language, resulting in 57 total artifacts responses. All participants were at least 18 years old; 28 shapes participants were female; 25 were female for artifacts.

3.1.3 Stimuli

Each experimental item consisted of a single image enclosed in a gray box on a computer monitor, a text box reading "*this item is...*," and three selection options: a gradable adjective (e.g. *tall*), its antonym (*short*), and *neither*. (10a-c) lists the full set of adjectives used in the experiment, which can be classified by their membership in the semantic classes of adjectives discussed in §1-2 above.

(10) a. *Relative:* long, tall, big, wide, thick, short, small, narrow, thin
b. *Maximum Absolute:* closed, empty, flat, full, plain, straight
c. *Minimum Absolute:* bent, bumpy, curved, open, spotted, striped

Shape images and adjective categories were drawn from the stimuli used in Aparicio et al. (2015), which consisted of artificially-constructed geometric shapes such as triangles, spirals, ovals, etc. that vary in size and color; see Figure 2a for an example. Visual properties of the shapes were selected to accommodate variation along the scalar dimensions of the adjectives associated with them; images associated with *full* and *empty*, for example, were always three-dimensional. Artifact-images had the same basic structure, the only difference being that the images were photographs of familiar objects such as stacks of books, candles, nails, shirts and so

forth; see Figure 2b for an example. Artifacts images corresponding to the adjectives in (10) were collected from various stock picture websites.

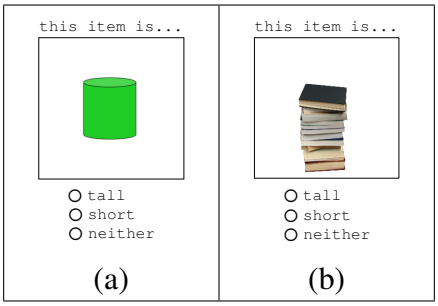

Figure 2: Typical items used to elicit judgments about gradable adjectives in shapes (a) and artifacts (b).

Individual items belonged to a larger set whose members differed only in the degree to which they exemplified some scalar dimension, as illustrated in Figure 3. From each scalar dimension, two scales can be defined, corresponding to a gradable adjective and its antonym; e.g. the height dimension provides a scale both for *tall* and for *short*. The items belonging to a common image set were thus assigned unique scale positions for two gradable adjectives: values determined by position in the ordering induced by each adjective. The greater an object's scale position, the better an exemplar it is of the adjective associated with the scale. For example, in the set of images used to assess judgments about the heights of cylinders shown in Figure 3a, the shortest cylinder was assigned the scale position 1 for assessing judgments about *tall*, but position 7 for assessing judgments about *short*; the second-shortest was assigned the value 2 for *tall* and 6 *short*; and so on.

Figure 3 illustrates sets of images that differ only in height, which were used to evaluate judgments about the relative adjectives *tall* and *short*. Figure 4 illustrates sets of images that differ only in stripedness, which were used to evaluate judgments about the absolute adjectives *striped* and *plain*.

As illustrated in Figure 3 and 4, the number of items per set differed across shapes and artifacts, with shape sets consisting of seven items and artifact sets consisting of five. This difference was necessary in part because of the materials used to create the experimental stimuli. Fine-grained manipulations of size, color, stripedness, etc., were easily performed on images of geometric shapes, and such manipulations were perceptually distinguishable. On the other hand, small gradations in scalar dimension were more difficult to achieve in images of real-world objects, and also tended to be less visually salient. Using 5- instead of 7-point scales for artifacts mitigated the risk that our perceptions of scalar distinctions might not correspond exactly to those of experimental participants.

Figure 3: Image sets forming scales for *tall* and *short* – shapes (a) and artifacts (b).

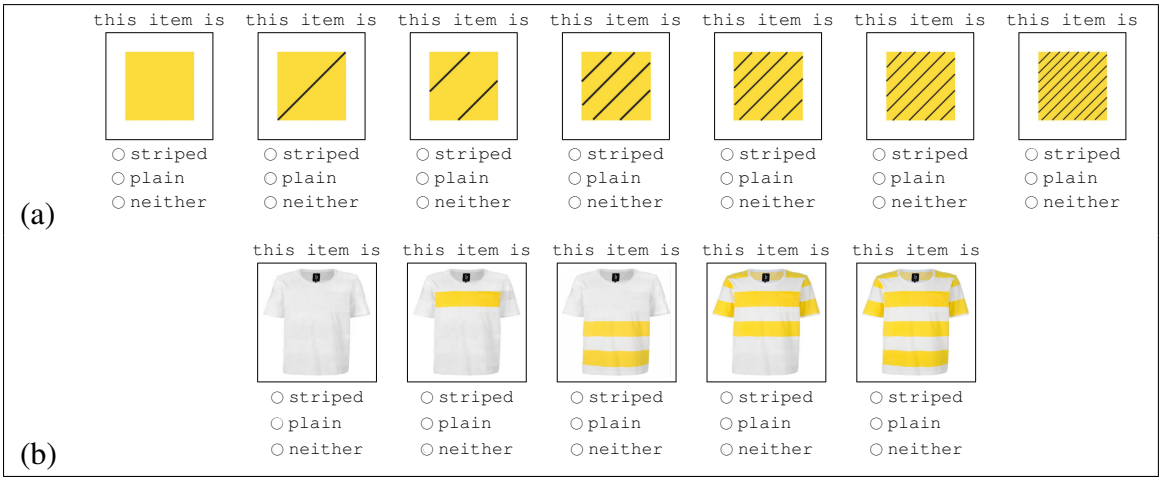

Figure 4: Image sets forming scales for *striped* and *plain* – shapes (a) and artifacts (b).

To facilitate comparison of judgments at different scale positions across shapes and artifacts, we created a general, five-point scale by superimposing the seven-point shapes scale onto the five-point artifacts scale. Scale positions for shapes were adjusted according to the following mapping: position 1 (the shapes scale minimum) was assigned to position 1 on the general scale; positions 2 and 3 were assigned to 2; position 4 (the shapes midpoint) was assigned to 3, the midpoint of the general scale; positions 5 and 6 were assigned to 4; and position 7 (the shapes scale maximum)

was assigned to 5, the maximum of the general scale. Artifacts' scale positions were maintained in their original form on the general scale.[5]

3.2 Procedure

3.2.1 Experiment administration

Experimental participants completed either 168 critical shapes trials (24 image sets $\times$ 7 items per set), or 120 artifacts trials (24 image sets $\times$ 5 items per set). Prior to beginning the experiment, subjects were told that they would see images of objects, and that for each object they should choose one of three responses to indicate whether they judged the object to have the property expressed by an adjective, its antonym, or neither.

Within each object type, approximately half of the participants in that version evaluated the items in isolated presentation, and the other half in grouped presentation. In the isolated presentation conditions, participants saw a single image on the screen, selected one of the three options for categorizing the object, and then hit a "next" button to move to the subsequent item. This process continued until every critical and filler item had been evaluated. Filler items asked subjects to indicate what color an object was (e.g. *red*, *blue* or *neither*). Images appeared in a different randomized order for each subject, and every subject saw every image with the same choices for adjectives.

Participants in the grouped presentation conditions saw a sequence of 40 screens containing scalar arrays of either seven shapes at a time, or five artifacts at a time. Arrays were presented in a different randomized order for each participant, in the manner specified above. Subjects were required to select a categorization option for every item in the group before they were able to hit "next" and proceed to a new set of images. For both artifact and shapes participants, 28 of the screens contained critical trials and 12 contained fillers in which they had to judge the color of seven (or five) images of similar but non-identical colors. As in isolated presentation, every subject evaluated every item exactly once.

Participation took no longer than twenty minutes. Upon completion of the experiment, participants had the opportunity to discuss the design and motivations of the study with the researcher who administered it (for educational purposes).

5 An alternative strategy for standardizing the shapes and artifacts scales could be to independently transform each scale to a set of z-scores centered around the mean of the respective original 5- or 7-point scale. Such a scale would have the advantage of maintaining a distinction between positions 2/3 and 5/6 for shapes, but would require that scale position be thought of as a continuous dimension with evenly-spaced points – a questionable assumption for the particular sets of images in this experiment (albeit probably justified in non-artificial scenarios).

3.2.2 Data transformation

Prior to analysis, we transformed response data into a binary format in the following way: where x is a trial with response options A (an adjective), $\bar{A}$ (its antonym), and $\emptyset$ (neither), the response $r(x)$ is coded as two distinct data points $r_A(x)$, and $r_{\bar{A}}(x)$ such that:

$$\text{if } r(x) = A, \quad \text{then } r_A(x) = \mathsf{true} \quad \text{and} \quad r_{\bar{A}}(x) = \mathsf{false};$$
$$\text{if } r(x) = \bar{A}, \quad \text{then } r_A(x) = \mathsf{false} \quad \text{and} \quad r_{\bar{A}}(x) = \mathsf{true};$$
$$\text{if } r(x) = \emptyset, \quad \text{then } r_A(x) = \mathsf{false} \quad \text{and} \quad r_{\bar{A}}(x) = \mathsf{false}.$$

The motivation for transforming the data in this fashion is twofold. First, each trial can conceptually be considered *two* judgments – one about A and one about $\bar{A}$ – under the (reasonable) assumption that A and $\bar{A}$ are mutually exclusive in this context.[6] Second, the binary-transformed *yes*/*no*-responses can be directly modeled as the dependent variable in logistic regression models (jaeger08; agresti02; barr13).

3.2.3 Data screening

We examined the response data by-subjects and by-items prior to analysis, to identify any anomolies arising from errors in the materials or participants who may not have taken the task seriously. We identified three items whose proportion of *yes*-responses by scale position deviated substantially from the others within the same adjective type: *long table*, *short table*, and *empty cube*. The source of these problems was incorrect assignment of images to scale positions in one case, and poorly selected or ambiguous images in the others. These items were removed from the dataset prior to analysis, yielding a total of 79 distinct items in the results dataset of record.

3.2.4 Statistical analysis

We analyzed the response data using logistic mixed-effects regression models with by-subject and by-adjective random intercepts.[7] Interactions between the fixed-effects predictors of scale position, adjective type, object type, and presentation type were investigated by fitting models to systematically defined subsets of the data,

6 *Neither* responses accounted for 15% of judgments, distributed across scale positions as would intuitively be expected: *neither* was most frequent at position 3, least frequent at positions 1 and 5, and in between at positions 2 and 4.

7 Full crossed random effects by participant and item were not possible in this case since object type and presentation type were both manipulated between-subjects; see barr13; jaeger08; barr08 for relevant discussion.

as described below. All models were fit in R v3.4.0, using `lme4::glmer()` with binomial family, logit link function, and BOBYQA optimization (lme4).

Because of the complexity of the study design and because of a strong expectation that effects of object type and presentation type would not be uniform across scale positions or adjective types, we used an iterative modeling strategy to assess the effects of object type and presentation type on *yes*/*no*-responses. We first fit a model to the full dataset with predictors for scale position, adjective type, object type, and all interaction terms involving them.[8] Interaction terms of interest were then investigated further by partitioning the data into subsets defined by the levels of adjective type (relative, minimum, maximum), and then refitting the model (without adjective type as a predictor) at each subset. Finally, the scale $\times$ object type interaction was broken down by further partitioning the adjective type subsets by scale position, and fitting even simpler models to each cell.

We assessed the impact of object type on log-odds of a *yes*-response at each subset of the data by performing a likelihood-ratio test (LRT) comparing the fit quality of the target model to that of a reduced (intercepts-only) model fit to the same data.[9] This test evaluates the hypothesis that the two models are no different from one another in terms of explanatory value, and a low p-value can be interpreted as evidence that the full model fits the data better than the reduced model (jaeger08). In other words, a low p-value from this test – at a particular scale point for a particular adjective type – suggests a significant effect of object type on probability of a *yes*-response.

3.3 Results

Proportion *yes*-responses for each adjective type and scale position are plotted in Figure 5 (averaged over all other grouping variables). Overall, response profiles for each semantic class of adjective had essentially the shape expected given their associated scale structure: maximum standard adjectives tend to be rejected until scale position 5; minimum standard adjectives tend to be accepted in all scale positions except for 1; and relative standard adjectives display a reasonably constant increase in proportion of *yes*-responses as scale position increases. These generalizations are consistent with the basic semantic characterizations in Section 1 above. Responses

8 Statistical procedures for evaluating object type were identical to those for presentation type, so here we describe only the former.

9 In all but two cases, the p-values computed with this procedure corresponded closely to those obtained by the Wald test for individual coefficients as specified in the default summary of a `glmer()` object. In the two diverging cases (minimum standard positions 1 and 5), proportion of *yes* responses across both levels of the predictor were very close to either 0 and 1, suggesting an artificial floor/ceiling effect as the source of the discrepancy (consonant with the very high (relative) standard errors for β_{artifact}). See Figure 8 for details.

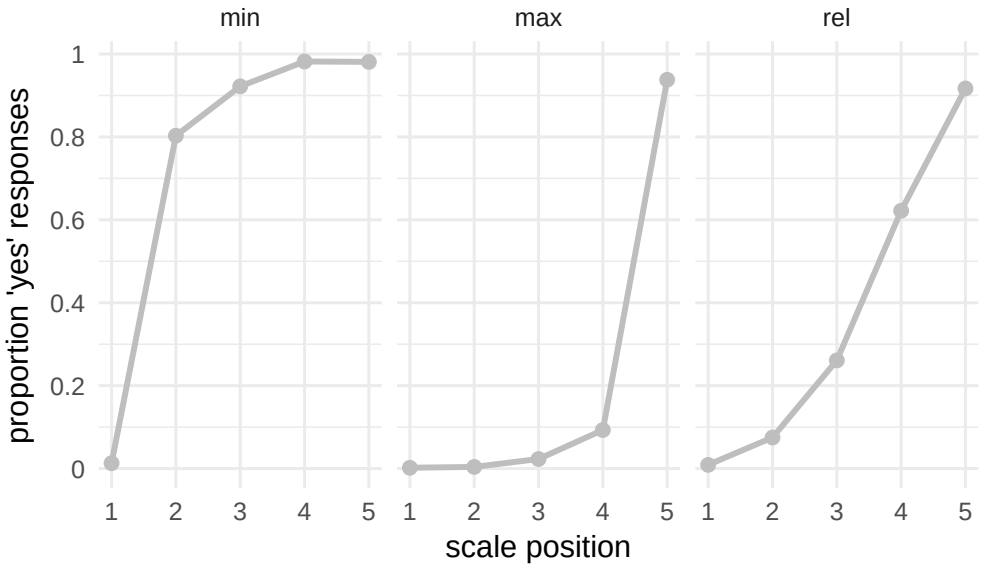

Figure 5: Proportion *yes*-responses by adjective type and scale position, averaged over participants and items.

for individual adjectives by scale position and object type are plotted in Figure 6. While there is certainly variation in judgments across adjectives, within each class the patterns are approximately the same.

We fit logistic mixed models with fixed effects for scale position, object type, and their interaction to each subset of the data defined by relative, maximum, and minimum standard adjectives. By-subject and by-adjective random intercepts were also included. The scale position-object type interaction was significant for maximum standard ($\beta = -3.0$ (std error .24); $\chi^2_1 = 189$, $p < .0001$) and minimum standard ($\beta = -.68$ (.14); $\chi^2_1 = 24$, $p < .0001$) adjectives, and was marginally so for relative standard ($\beta = -.12$ (.06); $\chi^2_1 = 3.7$, $p = .054$).

The same procedure for presentation type yielded significant interaction for relative ($\beta = -.29$ (.06); $\chi^2_1 = 23$, $p < .0001$) and minimum standard ($\beta = .44$ (.14); $\chi^2_1 = 10.4$, $p < .005$), but there was no meaningful interaction for maximum standard items ($\beta = .13$ (.21); $\chi^2_1 = .35$, $p = .55$).

Because of the heterogeneity of the scales (i.e. some positions are closer to an endpoint than others), finer-grained analysis is required to understand the nature of these interactions. The effects of object type and presentation type are broken down by individual scale points and discussed in more depth in the subsections that follow.

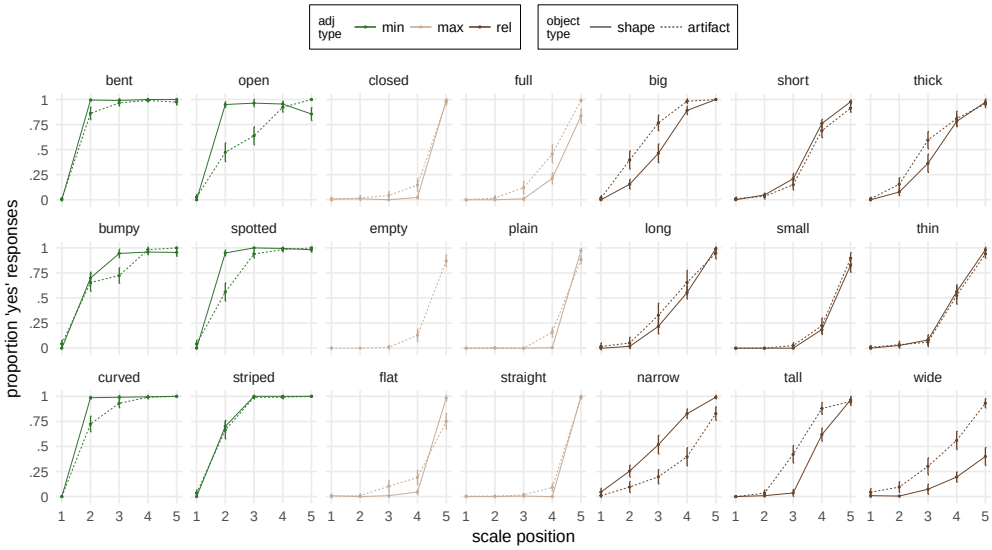

Figure 6: Proportion *yes*-responses for each adjective and scale position, in shape (solid) versus artifact (dashed) conditions. Adjectives are grouped by semantic class (minimum in green; maximum in beige; relative in brown). Error bars represent bootstrapped 95% confidence intervals computed at each point shown.

3.3.1 Analysis of object type

Response profiles for shape versus artifact object types are plotted in Figure 7. There was considerable variation in proportion of *yes*-responses across shape and artifact object types, with the most pronounced differences occurring for maximum standard adjectives at position 3 ($\beta = 2.5$ (std error .6); $\chi^2_1 = 31$, $p < .0001$) and position 4 ($\beta = 1.8$ (.24); $\chi^2_1 = 43$, $p < .0001$); for minimum standard adjectives at position 2 ($\beta = -1.5$ (.2); $\chi^2_1 = 46$, $p < .0001$) and position 3 ($\beta = -2.4$ (.33); $\chi^2_1 = 60$, $p < .0001$); and for relative standard adjectives at position 3 ($\beta = .57$ (.15); $\chi^2_1 = 15$, $p < .001$).

Table 8 summarizes the object-type regression results for each adjective type at each scale position. The general pattern among both classes of absolute gradable adjectives is that the shapes data coforms more closely to the categorical, endpoint-oriented semantics than do the artifacts data.[10]

10 It is worth noting that while we did not perform correction for multiple comparisons on the p-values presented here, all significant results other than relative-position 5 (object type), minimum-position 1 (object type), and minimum-position 2 (presentation type) would survive Bonferroni correction.

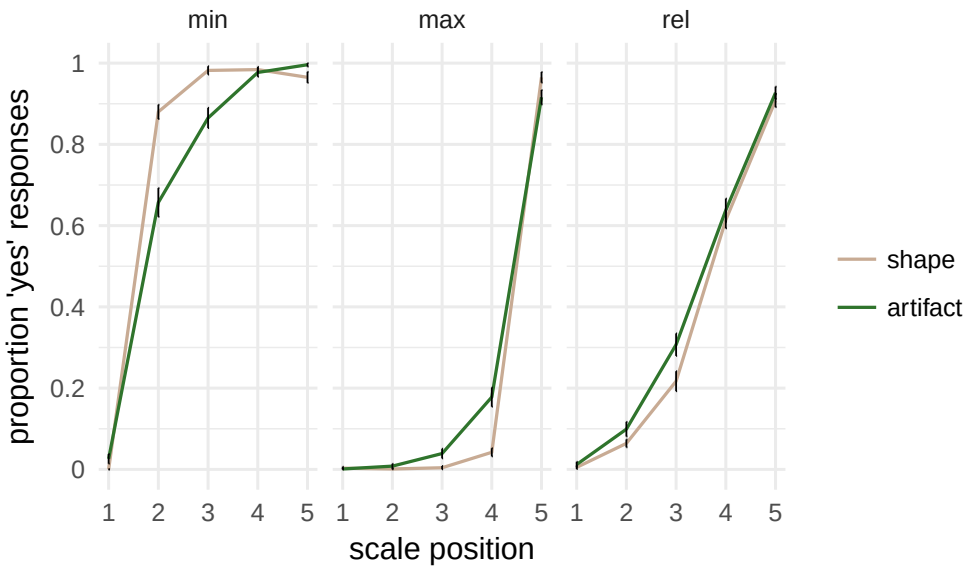

Figure 7: Proportion *yes*-responses for each adjective type and scale position, in shape (light/beige) versus artifact (dark/green) conditions. Black error bars represent bootstrapped 95% confidence intervals computed at each point in the plot (small intervals result from very large sample size and proportions close to zero or one).

	scale point	prop-*yes* (shp, art)	β_{artifact} (*se*)	$\chi^2(1)$	*p*-val	
max std.	1	.003, .001	-.88 (1.2)	.55	.46	
	2	.001, .008	1.7 (1.5)	1.2	.27	
	3	.004, .039	2.5 (.60)	30.8	$< .0001$	***
	4	.042, .178	1.8 (.24)	43.3	$< .0001$	***
	5	.965, .916	-1.2 (.37)	13.4	$< .001$	**
min std.	1	.000, .026	19 (44)	15	$< .001$	†
	2	.880, .657	-1.5 (.20)	46	$< .0001$	***
	3	.982, .865	-2.4 (.33)	60	$< .0001$	***
	4	.984, .977	-.55 (.51)	1.14	.29	
	5	.965, .996	3.0 (1.8)	4.3	.04	†
rel std.	1	.005, .012	.77 (.61)	1.6	.21	
	2	.064, .099	.39 (.19)	3.8	.05	·
	3	.217, .307	.57 (.15)	14.8	$< .001$	**
	4	.614, .638	.16 (.12)	1.85	.17	
	5	.908, .927	.71 (.27)	6.99	$< .01$	*

†Here LRT *p*-values were below .05 but Wald-test *p*-values on β_{artifact} within the models were not ($p = .66$ for 1; $p = .11$ for 5). In all other cases the two values were congruent; see fn9 for discussion.

Figure 8: Summary of regression results for the analysis of object type. Each row represents a single fit to a subset of the data defined by adjective type and scale position. Columns contain: proportion *yes*-responses (for shape, for artifact); model coefficient β_{artifact} for object type with standard error (note that shape is the reference value); the chi-square test statistic for the full vs reduced model; and *p*-value from LRT via nested model comparison with an intercepts-only model.

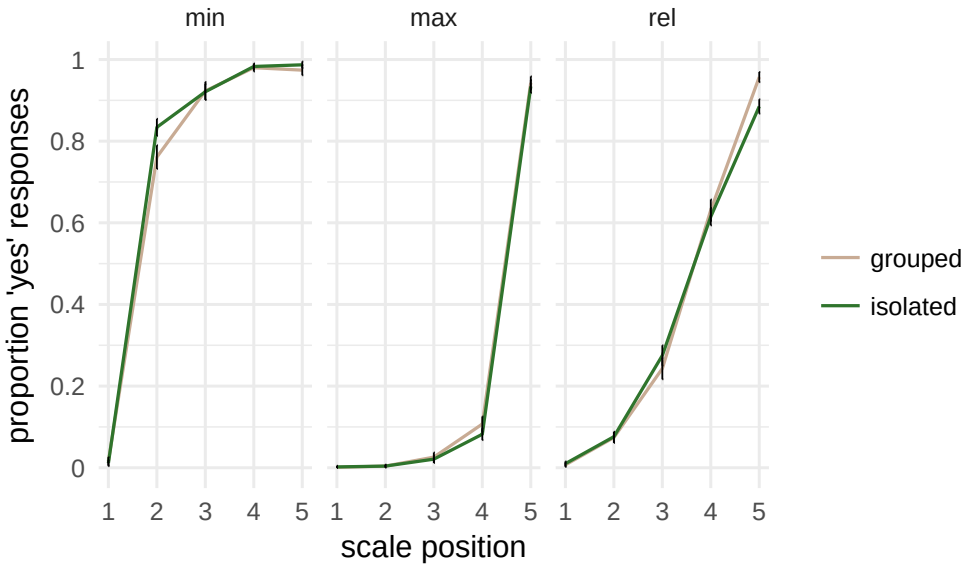

Figure 9: Proportion *yes*-responses by adjective type and scale position, in grouped (light/beige) versus isolated (dark/green) presentation. Black error bars represent bootstrapped 95% confidence intervals computed at each point in the plot.

3.3.2 Analysis of presentation type

Response profiles for isolated versus grouped presentation types are plotted in Figure 7. As can be seen from the figure, there was very little variation in proportion *yes*-responses across presentation types, the chief exceptions being position 2 for minimum standard adjectives, and position 5 for relative standard adjectives. These were the only cases in which a significant effect of presentation type was found ($\chi^2_1 = 4.5$, $p < .05$ with $\beta = .60$ (.28) for position 2 minimum; and $\chi^2_1 = 23.4$, $p < .0001$ with $\beta = -1.28$ (.25) for position 5 relative).

3.4 Discussion

To summarize, our experimental study demonstrated that in the case of impoverished prior knowledge about the distribution of objects on a scalar dimension (operationalized here as the "shapes" condition) absolute adjectives have more categorical, endpoint-oriented threholds compared to the case in which more prior knowledge can be assumed (the "artifacts" condition), in which they have more relaxed and variable thresholds. These results conform to the predictions of the scalar represen-

tation theory but not to the domain distribution theory, and so lend support to the hypothesis that the representational structure of closed scales gives rise to a convention for default minimal or maximal thresholds, with a move to more uncertain, distributionally determined thresholds only when justified by prior knowledge about the domain (or, potentially, other contextual factors).

Two aspects of this conclusion need to be qualified. First, the experimental design of the current study crucially assumes that participants have relatively poor prior knowledge about the distribution of geometric shapes with respect to those scalar properties tested in this study. But we are aware that there may exist individual variation both among participants and among the items we tested. Some participants may have had more experience with geometric shapes than others. And for some of our items, the particular scalar dimension we tested may not necessarily be unusual in specific contexts. For example, straight lines are fairly common objects to encounter in a math class. However, by and large, we think it is reasonable to assume that participants had significantly more previous experience with artifacts and their relevant scalar properties. We leave the issue of individual variation for future exploration.

A second assumption we made in evaluating the predictions for the distributional view is to (implicitly) adopt a uniform flat prior in the absence of substantive prior distributional knowledge (namely, for shapes). The uniform prior is relatively speaking an "objective" prior, and could be taken as a formal representation of ignorance. This is a reasonable and convenient assumption about the belief state of the language users when they encounter unfamiliar descriptions of objects. However, given that the choice of prior plays a critical role in deriving predictions for the distributional model, it remains an open question whether there are alternative "non-informative" or "objective" priors that the distributive model could also entertain, and whether such priors would make the distributional model a better fit to the observed data. We want to note that there is ongoing debate in the statistical modeling literature as to whether there exists truly objective priors. Instead of embarking on the pursuit of a true objective prior for situations that participants do not have much prior distributional world knowledge, one simple solution is to default to a prior justified by the scale structures themselves, as was done by (Qing & Franke 2014a: see section 2.2). This, of course, is exactly the choice made by the representational theory.

Even though our data from shapes is most compatible with a view that postulates a representational distinction between absolute and relative adjectives, we do not purport to invalidate the distributional approach altogether. In fact, the results from artifact images endorse the fundamental insight from the distributional model, i.e. that language users' prior world knowledge about how objects are distributed in a given comparison class has important impact on the decision process that classifies

whether an object has a certain adjectival property. If we take the difference in judgments between artifacts and shapes as indexing the effect size of world knowledge, we note that world knowledge has the strongest effect around the threshold area: across all three types of adjectives we tested, the largest changes from shapes to artifacts occurred at scale positions that are closest to the threshold region on each adjective scale. As shown by Figures 7 and 8), for maximum standard adjectives, participants were more likely to respond *yes* to a position 3 or 4 artifact image than they were for a shape image at the same region of the scale. And the effect for position 4 was larger than for position 3, suggesting that the effect becomes stronger as the scale approaches the threshold. For minimum standard adjectives, participants were *less* likely to respond *yes* at positions 2 and 3, the region just beyond the threshold. Similar to the maximum case, the effect is stronger closer to the threshold (here, stronger at position 2 than 3). For relative standard adjective, the effect is the strongest at scale position 3: participants were more likely to respond *yes* to a position 3 artifact image than they were for a shape image at the same region of the scale (albeit there was also a slight effect at positions 2 and possibly 5 as well).

Finally, we note that grouped versus isolated presentation type had barely any, or at most very subtle, effects on judgments. Significant effect of presentation type was only observed in two places. One place is relative standard, position 5 — grouped presentation yielded a higher proportion of *yes*-responses than did isolated; the other reliable difference in presentation type occurred in minimum standard at position 2, where grouped presentation was slightly less likely to yield a *yes* response (see section 3.3.2). We would caution against reading too much into these two observations, since presentation type yielded very little effect in general. The absence of a strong effect from presentation type is surprising, since one may have expected that grouped presentation provided participants with an explicit comparison class, which could result in different judgments from isolated presentations. The presentation effect was particularly expected for relative adjectives, which presumably were more sensitive to salient comparison class information. It is possible that for some of the relative adjectives we tested, especially the dimensional ones such as tall, big, etc, participants could have used other cues such as the height or size of their computer monitor screen to aid their judgments. But it is also possible that the lack of a strong presentation type effect may have a principled explanation that speaks to some deeper properties of how comparison class is constructed and evaluated for different types of adjectives. We leave this as an open question for future explorations.

4 Conclusions

This paper compared two different perspectives on the resolution of threshold uncertainty in the interpretation of gradable adjectives in context: one in which decisions

about thresholds are based entirely on probabilistic knowledge about how objects in an adjective's domain distribute relative to the scalar concept it encodes (domain distribution), and one in which the representational properties of the scale give rise to a convention for default endpoint-oriented standards when possible (scalar representation). These two views differ in their predictions about the interpretation of absolute adjectives in contexts involving impoverished distributional priors: the domain distribution hypothesis predicts more uncertainty about thresolds in such contexts, leading to more relative-like interpretations of absolute adjectives; the scalar representation hypothesis predicts more certainty thresholds at scalar endoints and more categorical absolute interpretations. Our experimental results indicate that although distributional information plays a key role in decisions about thresholds, scalar information takes priority in contexts involving impoverished distributional priors, in support of the scalar representation hypothesis.

References

Aparicio, Helena, Christopher Kennedy & Ming Xiang. 2015. Processing gradable adjectives in context: A visual world study. In Sarah D'Antonio, Mary Moroney & Carol Rose Little (eds.), *Proceedings of the 25th Semantics and Linguistic Theory conference*, vol. 25, 413–432. Linguistic Society of America and Cornell Linguistics Circle.

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.

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.

Kim, Christina, Christopher Kennedy & Ming Xiang. 2014. Shiftability and goal-dependence in gradable adjectives. Poster presented at the 88th Annual Meeting of the Linguistic Society of America, January 2-5, 2014, Minneapolis, MN.

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.

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Papers from the ESSLLI 2009 workshop on vagueness in communication* (cture Notes in Artificial Intelligence 6517), 151–168. Berlin: Springer.

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

Potts, Christopher. 2008. Interpretive Economy, Schelling Points, and evolutionary stability. Ms., University of Massachusetts, Amherst.

Qing, Ciyang & Michael Franke. 2014a. 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.

Qing, Ciyang & Michael Franke. 2014b. Meaning and use of gradable adjectives: Formal modeling meets empirical data. In *Proceedings of the annual meeting of the cognitive science society*, vol. 36, 1204–1209.

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.

Williamson, Timothy. 1992. Vagueness and ignorance. *Proceedings of the Aristotelian Society* 66. 145–162.

Word count: 10,265

Department of Linguistics
University of Chicago
Chicago, IL 60637

tjleffel@uchicago.edu
mxiang@uchicago.edu
ck@uchicago.edu