Proceedings of SALT 22: 166–187, 2012

Experimenting with degree *

Stephanie Solt
University of Amsterdam

Nicole Gotzner
Humboldt-Universität zu Berlin

Abstract Semantic theories differ in the role they assume for degrees in the interpretation of gradable adjectives, and in the assumptions they make about the nature of degrees and the structure of the scales they comprise. We report on two experiments investigating speakers' use of gradable adjectives across varying contexts, with the goal of gaining insight into the nature of the degree ontology underlying their semantics. We find that the truth conditions for the positive form must be stated in terms of degrees rather than rankings of individuals, and further that the relevant scale structure is one where distances between scale points are meaningful, and not an ordinal scale derived from an ordering relation on a comparison class. We also find no evidence that scale structure depends on the presence or absence of a corresponding system of numerical measures.

Keywords: gradable adjectives, comparison classes, scales, measurement levels, experimental semantics

1 Introduction

It is generally agreed that the interpretation of gradable adjectives such as *tall* and *dark* in some way makes reference to **degrees**. But what exactly degrees are, and which adjectival forms invoke them, remain open to debate. Focusing on what notion of degree, if any, underlies the semantics of gradable adjectives in their positive (unmodified) form, we take an experimental approach to investigating these questions. We show that by observing how speakers' use of such adjectives changes as the context is systematically manipulated, we can gain insights into the formal structures needed to represent their meaning.

As will be seen below, our results show that the interpretation of the positive form involves degrees organized into a scale in which the distance between scale

* We would like to express our appreciation to the participants at SALT 22, and to the conference reviewers, for much valuable feedback. Thanks also to the audiences at the ZAS Semantics Circle and at Linguistic Evidence 2012, where earlier versions of this work were presented, and especially to Manfred Krifka, for suggesting the basic idea for this research. All remaining errors are of course our own. Support for this work was provided by the European Science Foundation (ESF) and the Deutsche Forschungsgemeinschaft (DFG) under the auspices of the EUROCORES Programme LogICCC. The first author also acknowledges support from the NWO.

points is meaningful, and in particular are difficult to reconcile with proposals that scales are derived from ordering relations on sets of individuals.

The organization of the paper is the following. In Section 2 we introduce two leading semantic theories of gradability, as well as a recent proposal that represents an elaboration of one of them, and outline the predictions they make for speakers' use of gradable adjectives across different contexts (comparison classes). In Section 3 we present the results of two experiments where we asked subjects to do just this—apply gradable adjectives to comparison classes differing systematically in their composition—and assessed their behavior relative to these theoretically based predictions. Finally, in Section 4 we discuss the implications of our findings for the semantics of gradability, and suggest some possible extensions of our approach.

2 Theories of gradability and their predictions

2.1 Semantics of gradability

Two broad classes of semantic approaches to gradability can be distinguished, which have come to be known as **delineation**-based and **degree**-based. Their central components are described below.

Delineation The delineation approach is most closely associated with the work of Klein (1980, 1982). In Klein's theory, gradable adjectives denote context-dependent partial functions that induce a three-way partition of a comparison class C into a positive extension, a negative extension, and a so-called extension gap consisting of entities that fall in between the two. For example:

(1) $[\![tall]\!]^C = f\colon C \mapsto \{0,1\}$, where
$f(x) = 1$ iff x is tall
$f(x) = 0$ iff x is not tall
$f(x)$ is undefined otherwise

Here a comparison class is some subset of the universe of discourse that provides a frame of reference for the evaluation of the adjective (see below). To allow for the representation of measure phrases (e.g. *6 feet tall*), a system of degrees can be layered on top of the basic delineation approach, typically by defining a standard object and a concatenation operation on the domain. But importantly, this plays no direct role in the semantics of the positive construction.

Klein does not attempt to address what sort of heuristic speakers might employ in partitioning a given comparison class into the three specified subsets. Clearly, the tallest members of any comparison class must count (in that context) as *tall*,

and the shortest must count as *not tall*, but beyond this the truth conditions in (1) leave unspecified how the boundaries are to be drawn between *tall* and 'gap', and between 'gap' and *not tall*. Although degrees are not explicitly represented in (1), Klein's theory is consistent with a view in which this partitioning involves some psychological sense of degree, one which however is not part of the semantic formalism. But we could also hypothesize a strong version of the delineation approach in which no notion of degree at all is involved in the partitioning. It is this strong version in particular that we will consider.

Degree In contrast to delineation theories, the degree-based approach to gradability, associated with the work of Bartsch & Vennemann (1973); Cresswell (1977); von Stechow (1984); Bierwisch (1989); Heim (1985, 2000); Kennedy (1999, 2007), among others, takes gradable adjectives in their basic lexical semantics to encode relations between individuals and degrees on a scale. Formally, a scale is a triplet $S = \langle D, >, DIM \rangle$, where D is a set of degrees, $>$ is an ordering relation on that set, and DIM is a dimension of measurement. For example:

(2) $\quad [\![tall]\!] = \lambda d \lambda x.HEIGHT(x) \geq d$

The degree argument d in (2) is saturated or bound by degree morphology such as measure phrases or degree modifiers; in the case of the unmodified positive form, it is typically assumed that this role is played by a phonologically null 'positive' morpheme *pos*.

Degree theories differ substantially in their specifics. Of particular interest here are differences in the assumptions made about the nature of degrees and scales themselves (see Klein 1991 and more recently Sassoon 2010; van Rooij 2011; Lassiter 2011 for discussion of this point).

On one approach, which we might call the **abstract degree** theory, degrees are some sort of abstraction. For von Stechow, 'whatever they are, they are highly abstract objects' (von Stechow 1984: 47). Somewhat similarly, Kennedy and others describe degrees as 'abstract representations of measurement' (Kennedy 2007: 3). Although this point is not always made explicit, underlying this view is the assumption that degrees are a primitive component of the ontology, which have an existence independent of the entities whose measurements they encode. Just as the domain of individuals has been argued to have a structure which is linguistically relevant (cf. Link 1983), so too does the domain of degrees have potentially meaningful structure. Just what this structure is, and how this is linguistically reflected, is a matter for empirical study (see e.g. Kennedy & McNally 2005; Fox & Hackl 2006).

An alternate and more concrete view of degrees, which might be called the **derived degree** theory, holds that degrees and scales are derived from ordering relations on individuals. This view dates back to Cresswell, and can also be found

in work by Klein (1991); van Rooij (2011); Lassiter (2011) and, as will be seen below, Bale (2008, 2011). On this approach, we begin with a comparison relation *R* between individuals, such as a relation of 'taller than' or 'more beautiful than'. The equivalence classes under this relation become the degrees of the scale (where *x* and *y* are equivalent under *R* iff for all *z*, *R(x,z)* iff *R(y,z)*, and *R(z,x)* iff *R(z,y)*). A relation between degrees is then derived from the relation between individuals as follows: for equivalence classes (i.e. degrees) $\overline{a}$ and $\overline{b}$ containing individuals *a* and *b*, respectively, $\overline{a} > \overline{b}$ iff *R(a,b)*. In this way, a well-ordered scale is constructed. The derived degree theory thus bears an affinity to the delineation theory, in that relations between individuals are taken to be primary, and scales are derived from this rather than taken as primitive.

Derived degrees with numerical measures Bale (2008, 2011) proposes an extension to the derived degree approach which is particularly relevant to the present work because it makes testable predictions about speakers' use of different types of gradable adjectives across contexts. In Bale's theory, like that sketched out above, gradable adjectives are associated with relations, such as the relation '*x* is as beautiful as *y*', which are transitive and reflexive, and thus meet the criteria for a pre-order or quasiorder. Scales are constructed as above, by grouping entities into equivalence classes and deriving an ordering relation on the set of these classes. Finally, a measure function is defined that maps an individual to the equivalence class (i.e. degree) it belongs to.[1]

Bale's innovation is the following: for adjectives associated with a numerical system of measurement (e.g. height in meters, weight in kilos), measurements themselves (*2 meters*, *35 kilos*) enter into the underlying ordering relation as individuals. With some seemingly reasonable assumptions about which degrees are included, the result is a scale whose structure is isomorphic to that associated with the measurement system. Thus adjectives of this type come to be associated with scales with more structure than do those without measurement systems, such as *beautiful* and *intelligent*.

In broad terms, all of these approaches are able to account for the basic facts of gradability. What has not, however, been the subject of much discussion is that they are subtly different in terms of the possibilities they make available for how the truth conditions of the positive construction can be stated, and thus make subtly different

1 In Bale 2008, there is a further step in this process: namely, degrees on the primary scale are mapped to degrees on a universal scale isomorphic to the rational numbers between 0 and 1, which serves as the basis for all comparisons. This step is deemphasized in Bale 2011 and as such we do not consider it here. The points to be made below would also apply to a theory involving a universal scale; but see van Rooij 2011 for discussion of some potential issues with this proposed mapping.

predictions for how speakers' use of these adjectives will change across contexts. We discuss this below.

2.2 Comparison classes and standards

Regardless of which formal theory we choose to adopt, we must in some way account for the intuition that gradable adjectives in their positive form are interpreted relative to a comparison class that provides a frame of reference or standard of comparison. For example, the height John must have to establish the truth of (3a) and (3b) is clearly different, evidence of the semantic effect of the two different comparison classes (jockeys, basketball players) on the truth conditions.[2]

(3) a. John is tall for a jockey.

b. John is tall (even) for a basketball player.

While various work has focused on how, if at all, comparison classes enter into compositional semantics (Fults 2006; Kennedy 2007; Solt 2011; Bale 2011; among others), much less attention has been paid to how, precisely, the truth conditions for the positive construction should be stated relative to a given comparison class (a notable exception is Schmidt, Goodman, Barner & Tenenbaum 2009).

Here, there are various possibilities. For example, the truth conditions for *tall* relative to a comparison class C might be specified in any of the following ways:

(4) $[\![$ John is tall $]\!]^C$=1 iff ...

a. ... John is among the tallest $n\%$ of the Cs

b. ... $HEIGHT(john)$ is among the top $n\%$ of heights of Cs

c. ... $HEIGHT(john) > mean_{x \in C}(HEIGHT(x))$

Thus as one possibility, John might be considered tall if he is among the tallest n percent—say, the tallest third—of the members of C (per (4a)). Something of this general form is suggested by Bale (2011). Alternately, John might be considered tall if his degree of height falls within some specified subsegment of the range of heights corresponding to C—say, the top third of this range (per (4b)). Truth conditions of this form are proposed by Bale (2008). This formulation is also consistent with a proposal by Bierwisch (1989) that the standard of comparison relative to a

2 More accurately, these observations apply to what Kennedy (2007) calls relative gradable adjectives, a class that includes *tall, large, expensive, heavy* and many others, which are characterized by context-dependent standards. So-called absolute gradable adjectives such as *full/empty* and *wet/dry* do not exhibit the same degree of context sensitivity, and thus do not seem to invoke a comparison class in the same way (though see Toledo & Sassoon 2011 for a proposal that the difference is rather in the type of comparison class). The focus of the present paper is relative gradable adjectives.

comparison class C is fully determined by the range of degrees corresponding to C. Finally, the standard might be derived as an average over the comparison class, per (4c), a possibility suggested by Bartsch & Vennemann (1973) and von Stechow (1984), among others.

Of course, none of the formulas in (4) is, in the form given, adequate. The truth conditions stated here create sharp cutoffs between *tall* and *not tall*, and as such fail to capture the vagueness of gradable adjectives, which manifests itself in the existence of borderline cases (cf. Klein's extension gap). Any of these formulations could potentially be adapted to be vague rather than precise. For example, in (4a-b) vagueness might be linked to indeterminacy in setting the parameter n. In (4c), vagueness might be captured by taking the standard to be not the mean itself but rather a range around the mean whose width reflects, in some underspecified way, the degree of dispersion in the comparison class (cf. Solt 2011); such an approach also captures the intuition that to be counted as tall in the context of some set C, it is not sufficient merely to be taller than the average for a C (cf. Kennedy 2007).

However, since the primary focus of this paper is not the vagueness of gradable adjectives, we do not pursue any of these possibilities here. The simplistic formulas in (4) are sufficient to illustrate an important point: the semantics of the positive construction can be expressed in terms of a number of different measures. Most simply, (4a) is not expressed in terms of degrees at all, but rather on the basis of a **rank order** of individuals in C. That is, it does not require assigning something called degrees of height to individuals, but merely ranking the members of C by their heights, and then assigning some fixed proportion of individuals at the top of this ranking to the category *tall*. By contrast, both (4b) and (4c) are stated in terms of degrees of height; but the structure of the scales required to interpret these two formulas is different. Drawing on the typology of measurement levels introduced first by Stevens (1946) (see also Kranz, Luce, Suppes & Tversky 1971), we can see that (4b) can be interpreted relative to an ordinal scale, that is, a simple linearly ordered set of degrees lacking any notion of distance between scale points. We will call this **ordinal degree**. While an ordinal scale allows us to compare degrees ($a>b$), mathematical operations such as the addition of two degrees are not meaningful. As such, an ordinal scale is not sufficient to support truth conditions such as (4c), which require that we sum up the heights of all the members of C and divide the result by the cardinality of this set. For this, we require a more informative scale structure, one in which the distance between scale points is meaningful. In the typology of levels of measurement, this is an interval or ratio scale; to avoid being more specific, we will call this (for want of a better term) **measurement degree**.

Below we will see why these distinctions are relevant to the semantics of gradability more generally.

2.3 Predictions

The starting point for the present research is the observation that the different theories of gradability introduced above differ in the sort of degree structures they make available, and thus in the formulations of the truth conditions for the positive construction that they support.

The hypothesized strong version of the delineation theory—where the interpretation of the positive form does not involve even an implicit notion of degree—is compatible with truth conditions based on a rank ordering of individuals, as in (4a). But such a (hypothetical) theory would not be compatible with the degree-based truth conditions in (4b-c).

The derived degree theory supports the formulation in (4a), which could be restated in degree terms, as well as truth conditions based on ordinal degree, as for example (4b). But crucially, the above-described procedure for deriving scales creates an ordinal scale only, i.e. a linearly ordered set of equivalence classes as degrees. As such, it does not support truth conditions based on measurement degree, such as (4c), which requires a scale in which distances between scale points are meaningful. This limitation does not, however, hold for the abstract degree theory. Such a theory takes the scales underlying natural language meaning to have an independent existence outside of any set of individuals to be measured, and in particular involves no *a priori* commitment to the structure of these scales. If scales are assumed to be a reflection of our ability to perceive or judge the magnitude of an entity's height, beauty, etc., then their structure must be at least as informative as our perceptions and judgments themselves—potentially including a representation of how far apart two degrees are. Thus this type of theory is compatible with truth conditions based on measurement degree, such as (4c).

Finally, Bale's version of the derived degree theory allows for a possible difference in behavior between adjectives that are associated with a numerical measurement system (e.g. *tall*) and those that are not (e.g. *beautiful*). For the latter, only ordinal degree is available, just as in the basic derived degree theory. But for the former, the inclusion of measurements (e.g. *1 cm*, *2 cm*, *3 cm*, ...) in the domain of the underlying ordering relation on individuals creates a richer scale structure, one with units at regular increments corresponding to those of the measurement system; that is, it creates measurement degrees. Thus for this class of adjectives, there are more possibilities for how the truth conditions for the positive construction can be stated.

This pattern of compatibilities is summarized in Table 1. In our experimental work, we test these predictions by investigating speakers use of gradable adjectives across comparison classes varying in their composition.

	Delineation (strong)	Derived Degree	Abstract Degree	Derived Degree with Numerical Meas.
Rank order	YES	YES	YES	YES
Ordinal degree	NO	YES	YES	YES
Measurement degree	NO	NO	YES	only adjectives with measurement systems

Table 1 Compatibility of theories and measures.

3 Experiments

The experiments reported here were designed to investigate speakers' interpretations of gradable adjectives relative to the alternative predictions outlined above. Specifically, we manipulated the statistical distribution of elements in a comparison class and measured how this manipulation affected which items speakers considered as *tall, dark*, etc., following a method developed by Barner & Snedeker (2008) and Schmidt et al. (2009). Our overall goal was to determine which measures are required to describe speakers' judgments, and on this basis to assess the empirical adequacy of the various theories of gradability discussed above. We do not in this research attempt to arrive at the ultimate formulation of the truth conditions of the positive construction, but rather to identify what structures the model must include in order to state the truth conditions.

First, we explore whether the semantics of gradable adjectives in their positive form can be expressed in terms of rankings of individuals, or whether it is necessary to introduce degrees. Secondly, we investigate what sort of scale structure is needed, if the semantics of the positive form reference degrees.

3.1 Experiment 1

In our first experiment, we focus on the question of whether speakers' judgments are based on a simple rank order of individuals in a comparison class, or whether they rely on some notion of degree.

3.1.1 Methodology

Participants The study was executed online via the Amazon MTurk platform (see Sprouse 2011 on the validity of MTurk for the collection of linguistic judgments). The survey was only displayed to subjects with U.S. IP addresses and those were further screened for native language. A total of 208 subjects took part, receiving 15 cents for participation. The data of 10 subjects were excluded from the analysis

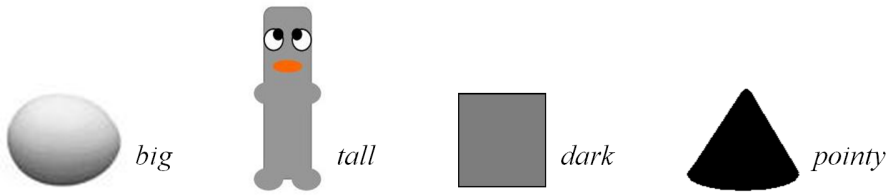

Figure 1 Experiment 1: adjectives and pictures.

because they indicated a native language other than English. Another 4 subjects were excluded due to inconsistent answers (n=1) or because they participated in more than one version of the survey (n=3). This left 194 native speakers of English for analysis (124 female; mean age 35.7).

Materials and procedure The stimuli for this experiment were based on four adjectives: *big*, *tall*, *dark* and *pointy*. Each was paired with arrays of 36 pictures spanning 11 'degrees' of size/height/etc., intended to represent comparison classes (respectively: eggs varying in size, cartoon characters varying in height, gray squares varying in RGB value, triangular shapes varying in angle and bluntness; see Figure 1). The adjectives were chosen such that two had corresponding numerical measurement systems (*big* - size, e.g in cm^2; *tall* - height, e.g. in cm) while two did not (*dark* and *pointy*).[3] The picture stimuli were selected in a way to encourage participants to base their judgments on the given stimuli set, not some prior notion of what *tall*, etc. might be; hence items were shown without context, and we used for the most part cartoon characters and abstract shapes instead of e.g. pictures of people.

We constructed 4 versions of each stimulus set that differed in how the 36 items (pictures) were distributed over degrees: a baseline peaked distribution (largest # of items of middle degrees; fewer high/low degrees), a left skewed and a right skewed distribution, and a moved distribution, featuring the same shape as the baseline distribution but shifted 3 degrees to overall greater sizes/heights/etc. Figure 2 shows the # of items per degree in the 4 distributions, and an example stimulus array.

The survey was developed in four versions, such that each respondent saw each of the 4 adjectives in 1 of the 4 distributions, in a Latin square design. At the beginning of the survey, we collected basic demographic information (native language, age,

3 A reader might object that these dimensions could be given numerical measures, or have such measures in specialized domains. For example, brightness/darkness can be measured relative to the Munsell scale. What we mean here is that in everyday speech we have no widely used numerical units of darkness or pointiness. We acknowledge, however, that the distinction is a blurry one.

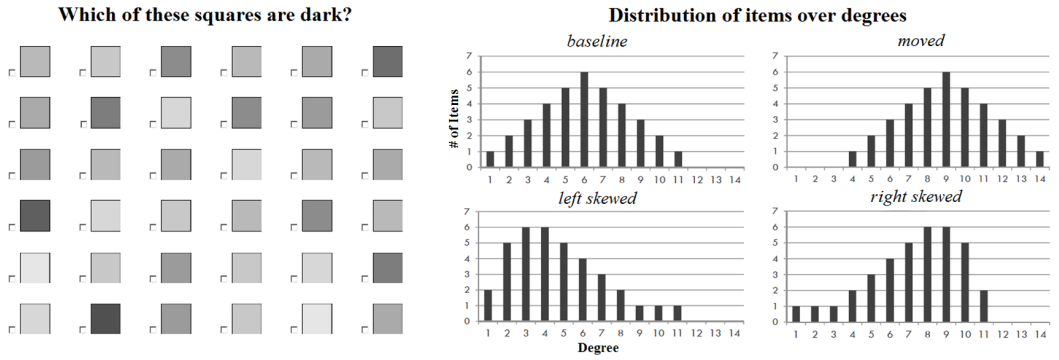

Figure 2 Experiment 1: sample array and distributions (# items per degree).

gender and education level) from the participants and we gave a description of the survey. The instructions told the participants that they would see 4 sets of pictures and that they should check the pictures that matched the description displayed above the pictures. They were also told that they could check as many or as few pictures as they liked. At the end of the experiment, participants could leave a comment, and from the comments none of the subjects seemed to have been aware of the experimental manipulation.

Predictions The rationale behind the experimental manipulation was that the underlying interpretation of the adjectives will be reflected in different classification patterns across the experimental conditions (i.e. distributions) and this will allow us to explore which approach most closely predicts speakers' behavior.

We formulate the following predictions:

- If speakers' judgments are based on a rank order of individuals (per (4a)), the number of items called *tall/dark*/etc. should be the same across all four distributions. For example, if the tallest 1/3 of individuals in any given comparison class are considered tall, then speakers should consistently check 12 out of the 36 pictures in each array.
- If speakers' judgments are based purely on an (ordinal) notion of degree (per (4b)) the 'cutoff point' should be the same in the baseline, left skewed and right skewed distributions, and higher in the moved distribution (where by cutoff point we mean the smallest item called big, etc.). For example, if *tall* corresponds to the top 1/3 of the ordinal degrees for a comparison class, then we expect the cutoff point to be roughly degree 8 in the baseline, left skewed and right skewed distributions, and 11 in the moved distribution.

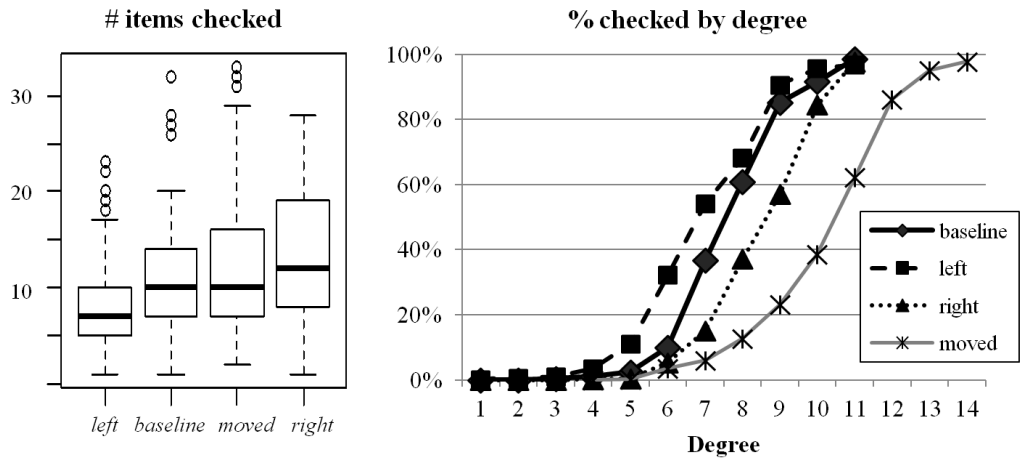

Figure 3 Results of Experiment 1: averages across adjectives and subjects.

3.1.2 Results

The answers of the participants were coded according to the two measures described above. The number of items checked was calculated for each participant in a given condition. For each participant, the cutoff point was determined as the first degree for which items were checked, without there being any higher degree for which no items were checked.

As seen in Figure 3, we found differences between distributions both in the total number of items checked and in the average proportion of items checked by degree (note that the cutoff point essentially provides a summary measure of the location of the curves in Figure 3). We submitted # items checked to a linear mixed effects model with Distribution and Adjective as fixed factors and Subject as random factor. Overall the effect of distribution on # items checked was significant ($F(3,760)=51.68$; $p<0.0001$). This indicates that we can falsify our first prediction that # items checked is not affected by Distribution. Table 3 in the Appendix summarizes the details of the mixed effects analysis. The model revealed a difference between the baseline and the left skewed distribution ($p<0.05$). The individual comparisons of the baseline and the right skewed distribution were not significant, nor was that between baseline and moved distribution significant. Note, however, that the moved distribution had exactly the same shape as the baseline distribution; therefore, this result is expected. There were also significant effects for the adjectives *tall* and *pointy*, which had a higher # items checked than did *big*. Finally, there was a significant interaction between the adjective *pointy* and the right skewed ($p<0.0001$) and moved ($p<0.0001$)

distributions, reflecting that more items were classified as *pointy* in these distributions compared to the baseline, while there was no such effect for the other adjectives.

The same model was fit to the cutoff point and it revealed reliable effects of the distribution in all comparisons (baseline vs. left skewed $p<0.05$; vs. right skewed $p<0.001$; vs. moved $p<0.0001$), falsifying the prediction that ordinal degree alone is sufficient to predict speakers' judgments. See Table 4 for details of the model. Again, there was a difference between *tall* and *big* (lower cutoff for *tall*), as well as an interaction between *pointy* and the right skewed and moved distributions ($p<0.01$ and $p<0.0001$ respectively), reflecting the fact that the cutoff point in these distributions differed less from the baseline than it did for the other adjectives.

3.1.3 Discussion

The results of our first experiment show first of all that the interpretation of gradable adjectives in their positive form cannot be stated in terms of rankings of individuals, as in the formula in (4a). *Tall*, for example, does not mean 'in the tallest *n%* of the relevant comparison class'. Rather, we need to invoke a notion of degree of height, size, shade, etc.

These findings do not yet provide strong evidence as to which notion of degree is necessary, and in particular do not resolve the question of whether an ordinal scale derived from an ordering on a comparison class (per the derived degree theory) is sufficient to model speakers' judgments. Very simple truth conditions based on ordinal degree alone (i.e. (4b)) are not compatible with these findings; thus *tall*, for example, does not correspond to some fixed subset of the set of degrees of height representing the comparison class.[4] However, we cannot rule out the possibility that the truth conditions involve some more complex formula based on rank order and ordinal degree. That is, we do not have conclusive evidence for the need to introduce measurement degree. The reason is that the 14 ordinal degrees of height, etc. that our stimuli were based on represented 14 evenly spaced steps along the relevant dimensions. That is, in our stimuli sets, ordinal degree and measurement degree were highly correlated, and thus we cannot determine which of the two is the crucial one. In Experiment 2 we address this by creating stimuli where this correlation is not present, allowing us to tease apart the two notions of degree.

A further observation that can be made about our findings from this stage is that we do not observe a qualitative difference in the behavior of adjectives with and without numerical measurement systems, as might be expected under the theory

4 This conclusion is somewhat at odds with the findings of Schmidt et al. (2009), who conclude that such a formula is one of the two most plausible candidates for the truth conditions of *tall*. We hypothesize that the nature of the distributions we tested, in contrast to those included in their research, allows us to eliminate a further possibility that was left open by that study.

developed by Bale (2008, 2011). While we did find between-adjective differences, they did not pattern along the lines of 'numerical' vs. 'non-numerical'. In its sensitivity to manipulations of the context, non-numerical *dark* behaved comparably to numerical *big* and *tall* (as evidenced by the lack of interactions). The behavior of non-numerical *pointy* did diverge from that of the other three adjectives. But here, judgments of *pointy* were found to be less rather than more context sensitive than those observed for the other three. Thus, participants seemed able to rely on some absolute threshold degree of pointiness that was independent of the structure of the contextually provided comparison class. Whether this different pattern is due to the semantics of *pointy* itself, the nature of our stimuli, or some combination of the two, is not certain (and as such we exclude *pointy* from the next stage of research). But in any case these findings tend to contradict rather than support a view in which the only notion of degree available for non-numerical adjectives is a scale constructed from a comparison class. We investigate the numerical/non-numerical distinction further in the next experiment.

3.2 Experiment 2

Results of the first experiment indicate that the truth conditions of sentences with gradable adjectives must be stated in terms of degrees; but these findings do not fully address what notion of degree is relevant. Specifically, is an ordinal scale constructed from a ranking of comparison class members sufficient (per the derived degree theory), or is it necessary to assume a more informative scale structure independent of the comparison class (as is possible under the abstract degree theory)? Experiment 2 was designed to answer this question.

3.2.1 Methodology

Participants In total, 200 subjects were recruited on the Amazon MTurk platform, with the study again restricted to U.S. IP addresses. 18 were excluded based on the native language criterion, 2 subjects gave inconsistent answers and another 10 subjects had already taken one of the other versions of the survey. 170 native speakers of English were included in the final analysis (111 female; mean age 30.4) and they again received 15 cents for participation.

Materials and procedure In Experiment 2, the adjectives *big*, *tall*, and *dark* were again paired with arrays containing 36 pictures (as before, eggs, cartoon characters and gray squares, respectively). For each adjective, 3 distributions were created (per Figure 4); each included items of 6 ordinal degrees of size/height/shade, but in absolute terms (i.e. in terms of measurement degree) featured a 'gap' somewhere

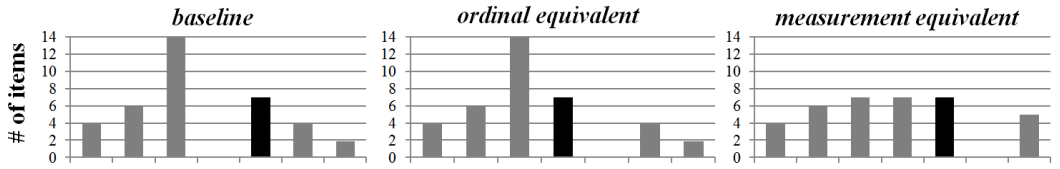

Figure 4 Experiment 2: distributions.

in the distribution. In the baseline distribution, the gap was between the 3rd and 4th ordinal degrees (more specifically, the 6 degrees here correspond to degrees 1,3,5,9,11 and 13 from Experiment 1). In the ordinal-equivalent distribution, the gap was between the 4th and 5th ordinal degrees (the corresponding Exp. 1 degrees being 1,3,5,7,11,13), while in the measurement-equivalent distribution, the gap was between the 5th and 6th ordinal degrees (corresponding Exp. 1 degrees 1,3,5,7,9,13). These three distributions were further designed to provide a 'target degree' that served as the basis for comparison across distributions (shown in black in Figure 4). In the baseline distribution, the target degree was ordinal degree 4 out of 6. In the ordinal-equivalent distribution, the target degree was the same ordinal degree as the baseline target degree (i.e. 4 out of 6) but lower in absolute (measurement) degree. Conversely, in the measurement-equivalent distribution, the target degree was the same in absolute (measurement) degree as the baseline target degree, but higher in ordinal degree (5 out of 6). The number of items of the target degree, and the number of items greater in degree, was held constant across distributions.

The procedure was the same as for Experiment 1. The study was executed in 3 versions, such that each subject responded to all 3 adjectives and distributions in a Latin square design.

Predictions Our predictions focus on the proportion of items of the target degree that are checked. Specifically:

- If speakers' judgments are based on ordinal degree derived from the comparison class, then the % target degree items checked in the baseline and ordinal-equivalent distributions will be equal.
- If the availability of measurement degree is dependent on an adjective's having a corresponding numerical system of measurement, then a difference between baseline and ordinal-equivalent distributions may be observed for the adjectives *big* and *tall*, but (crucially) not for the adjective *dark*.

	baseline	ordinal equivalent	measurement equivalent
big	44	1	14
tall	85	11	45
dark	41	10	32

Table 2 Results of Experiment 2: % target items checked.

3.2.2 Results

The results of Experiment 2 are displayed in Table 2. The proportion of target degree items checked was submitted to a linear mixed effects model with Distribution and Adjective as fixed effects and Subject as random factor. The model found a significant difference between baseline and ordinal-equivalent distributions ($p<0.0001$), disconfirming the first prediction. A significant difference was also found between the baseline and measurement-equivalent distribution ($p<0.0001$), further indicating that speakers' judgments are not based on degree alone. A detailed summary of the model can be found in Table 5 in the Appendix.

The model further revealed differences between the three adjectives tested, as well as interactions between adjectives and distributions; in particular, the % checked was higher for *tall* vs. the other two adjectives, consistent with the lower cutoff for *tall* found in Experiment 1 (see Table 5 for specifics). However, as seen in Table 2, the difference between baseline and ordinal-equivalent distributions was also found for the adjective *dark*; the significance of this difference is confirmed by an ANOVA ($F(2,140)=7.46$; $p<0.001$). Thus the second prediction is also disconfirmed, in that effects are noted not only for numerical *big* and *tall*, but also for non-numerical *dark*.

3.2.3 Discussion

From the perspective of the derived degree theory, the stimuli sets for the baseline and ordinal-equivalent conditions are indistinguishable: both give rise to identical 6-point ordinal scales, and both feature the same distribution of items over (ordinal) degrees. As such, the prediction of this theory is that they will elicit the same responses from subjects. What we found was different from this—dramatically so. Specifically, a significantly lower proportion of target items was checked in the ordinal-equivalent vs. baseline distribution, reflecting the fact that the items were lower in measurement degree. These findings are not accounted for by the derived degree theory, but rather require us to assume that respondents have access to a notion of degree independent of the structure of the comparison class, one that recognizes the magnitude of differences in degree between entities.

Nor did we find the difference between numerical and non-numerical adjectives predicted by Bale's mixed theory. Specifically, responses to the baseline and ordinal-equivalent distributions patterned differently even for the adjective *dark*, which corresponds to a dimension for which we have no commonly used numerical measure. Thus we find no evidence that the availability of measurement degrees is contingent on the existence of a numerical system of measurement.

The difference found between baseline and measurement-equivalent distributions is less relevant to our main research questions, but further supports the finding from Experiment 1 that degree alone is not sufficient to account for speakers' judgments; distribution of items over degrees also matters.

4 Conclusions and general discussion

The results of our two experiments show that the meaning of gradable adjectives in the positive form cannot be stated in terms of rank orderings of individuals; rather, we need degrees. Furthermore, the relevant scale structure is one in which the distance between scale points is meaningful, and thus cannot be an ordinal scale constructed from an ordering on a comparison class. And this is the case not only for adjectives associated with dimensions for which we have common numerical measures, but also for those without corresponding measurement systems.

Before we discuss the implications of these findings in more depth, we would like to address one potential objection. There is an assumption underlying our experimental approach, namely that it is the array of stimuli pictures shown to respondents that provides the domain or comparison class responsible for contextual effects. That is, we have assumed that respondents would calculate the truth conditions of the positive form based on this set alone, and more basically, with respect to the derived degree theory, that it would be this set alone that would serve as the basis for constructing the scale itself that is used in these calculations. This need not be the case. For example, respondents might have inferred that the stimuli items shown were merely sampled from a broader domain with a potentially different distribution. Or, despite our attempts to decontextualize the stimuli, respondents might have brought some sort of real world knowledge to bear on the task (cf. Tribushinina 2011 for evidence of this kind of effect).

While we cannot entirely discount these possibilities, we have reason to think that they did not contribute to our findings in a significant way. Most basically, respondents' behavior varied systematically in response to our manipulations of the stimulus set, evidence that this set influenced whatever calculation of truth conditions they performed. More direct evidence comes from the results of Experiment 1, specifically the comparison between the baseline and moved distributions. Recall that the moved distribution had the same shape as the baseline distribution, but

shifted 3 'steps' to overall greater degrees of height, size, etc. Correspondingly, for three of the four adjectives tested, the average cutoff point was almost exactly 3 degrees higher in the moved than the baseline distribution. The implication is that judgments of which items were big, tall or dark was based purely on the stimulus set given, not external considerations (the anomalous behavior of *pointy* perhaps then reflecting the intrusion of real world standards). On this basis we conclude that we are justified in assuming the comparison class can be equated with the stimulus set displayed. We nonetheless believe that this issue merits further investigation. One possibility would be to test stimuli sentences with an overt *for*-phrase (e.g. *tall for a Martian*), as such phrases are known to determine the comparison class.[5]

Having discussed this potential issue, let us turn to the implications of our results. In demonstrating that the interpretation of the positive form is based on degrees of the relevant dimension, our findings are most readily compatible with a degree-based theory of gradability, as in (2). But we have not necessarily demonstrated that degrees must be part of the formal representation of the positive form; that is, we have not shown that a lexical entry like Klein's (1) is inadequate. While our results are inconsistent with the hypothesized strong version of the delineation theory, we have not ruled out a version in which some psychological sense of degree (e.g. degree of height) is part of the conceptual basis on which we are able to categorize a given set into (for example) the tall and not tall members. We suspect that this question cannot be fully resolved by experimental approaches. We will point out, however, that the truth conditions for some adjectival constructions (notably measure phrases) must be represented in terms of degrees, and secondly that, according to our findings, the positive form is interpreted with reference to degrees. The most parsimonious account would thus seem to be one in which the semantics of all adjectival forms, including the positive, are explicitly based on degrees.

With regards to the nature of the degree structures underlying the interpretation of gradable adjectives, our results are clear. The truth conditions for the positive form cannot be calculated on the basis of an ordinal scale constructed from a ranking of comparison class members. Rather, speakers have access to a notion of measurement degree that is independent of the structure of a given set of individuals, as is allowed under the abstract degree theory. The procedure assumed by the derived degree theory is formally solid; but we find that a scale constructed in this way does not form the basis of speakers' judgments of what is big, tall, and so forth. Furthermore, we see no evidence that measurement degrees are derived from elements of a measurement system; if anything, we suspect the reverse is true.

It is important to note that the derived degree theory can be further elaborated to produce a scale with the required structure. What is needed is to define a concatena-

5 We thank an anonymous SALT reviewer for this suggestion.

tion operation $\circ$ on individuals of the domain, and require that for any a and b, the measure of $a \circ b$ is the sum of the measures of a and b; the result is a ratio scale (cf. Kranz et al. 1971). For height, for example, concatenation corresponds to stacking entities end on end: the height of two stacked entities is the sum of their individual heights. But while this is conceptually straightforward in the case of dimensional adjectives such as *tall*, it is far less obvious what the concatenation operation would be in the case of adjectives such as *dark*, *warm* or *beautiful*. Thus further work would be needed to determine whether this is a viable account of the scales speakers rely on in interpreting such adjectives.

We see considerable opportunity to extend the experimental approach followed in this research. At this point, our conclusions are based on a rather limited sample of adjectives, and in Experiment 2 in particular we had only one adjective representing the 'non-numerical' class. We hope to expand our empirical base in future research. Beyond this, while we found no role for ordinal scales for the adjectives we tested, perhaps other adjective classes, or other forms than the positive, have semantics based on such scales. Particularly interesting are evaluative adjectives like *beautiful* and *intelligent*, and emotion words like *happy*, as these provided the original impetus for the derived degree approach (Cresswell 1977), and there is little *prima facie* evidence that their underlying scales are anything more than ordinal in level (though see Sassoon 2010 for an alternate view). Here, there are some methodological challenges: it is quite difficult to find examples of adjectives without common numerical units, but for which the stimuli can nonetheless be manipulated in regular increments (as for instance our *dark* stimuli involved regular increments of RGB value). Some otherwise good candidates must be classified as absolute rather than relative gradable adjectives (e.g. *rough/smooth*), while others do not lend themselves to evaluation in an online context (e.g. *hard/soft*). Adjectives of the *happy* and *beautiful* sort are especially challenging. But if appropriate methodologies can be developed, experimental work has the potential to shed new light on the scale structure underlying the semantics of a wider range of gradable adjectives.

We conclude with some broader remarks. Our findings are unexpected from the perspective of one formal approach to gradability; but from another perspective they are entirely unsurprising. A large body of work from the field of psychophysics has shown that subjects are able to make ratio-level judgments on a wide variety of dimensions, including not only easily measured ones such as length and weight, but also those for which we have no commonly used measurement units, such as loudness, saltiness, roughness, pain and many others (Stevens 1975 and subsequent work). Findings of this sort have the potential to inform semantic theory, by providing a psychological perspective into the nature of the degree ontology that underlies the semantics of gradability. The structure of this domain remains a topic for empirical study; but we believe it must include something like measurement degree.

Appendix

In this appendix, we provide the details of the linear mixed models reported in Section 3 (note that non-significant interactions are not displayed).

Fixed Effect	Estimate	SE	t-value	pMCMC
(Intercept)	9.08333	0.74142	12.251	
left skewed	-2.18972	1.05409	-2.077	0.0322
right skewed	-0.28741	1.04317	-0.276	0.7736
moved	-0.44333	1.03799	-0.427	0.6736
dark	0.01667	1.03799	0.016	0.9706
pointy	2.40603	1.05409	2.283	0.0178
tall	3.01871	1.04317	2.894	0.0042
pointy:right skewed	8.01805	1.47557	5.434	0.0000
pointy:moved	8.87064	1.59401	5.565	0.0000

Table 3 Experiment 1: # Items checked.

Fixed Effect	Estimate	SE	t-value	pMCMC
(Intercept)	7.62500	0.20361	37.4	
left skewed	-0.62500	0.28947	-2.16	0.0312
right skewed	1.08929	0.28647	3.80	0.0002
moved	2.95500	0.28505	10.37	0.0000
dark	0.01500	0.28505	0.05	0.9560
pointy	-0.41223	0.28947	-1.42	0.1458
tall	-0.68622	0.28647	-2.40	0.0164
pointy:right skewed	-1.12205	0.40521	-2.77	0.0058
pointy:moved	-1.85527	0.43719	-4.24	0.0000

Table 4 Experiment 1: Cutoff point.

Fixed Effect	Estimate	SE	t-value	pMCMC
(Intercept)	0.43848	0.04811	9.113	
ordinal equivalent	-0.42804	0.06211	-6.891	0.0000
measurement equivalent	-0.30339	0.06520	-4.653	0.0000
dark	-0.03066	0.06520	-0.470	0.6298
tall	0.41045	0.06229	6.590	0.0001
dark:measurement equivalent	0.20691	0.09799	2.111	0.0214
tall:ordinal equivalent	-0.30925	0.09023	-3.428	0.0002

Table 5 Experiment 2: % target degree items checked.

References

Bale, Alan Clinton. 2008. A universal scale of comparison. *Linguistics and Philosophy* 31(1). 1–55. doi:10.1007/s10988-008-9028-z.

Bale, Alan Clinton. 2011. Scales and comparison classes. *Natural Language Semantics* 19(2). 169–190. doi:10.1007/s11050-010-9068-0.

Barner, David & Jesse Snedeker. 2008. Compositionality and statistics in adjective acquisition: 4-year-olds interpret *tall* and *short* based on the size distributions of novel noun referents. *Child Development* 79(3). 594–608. doi:10.1007/s10988-006-9004-4.

Bartsch, Renate & Theo Vennemann. 1973. *Semantic Structures: A Study in the Relation between Syntax and Semantics*. Frankfurt: Athaenum Verlag.

Bierwisch, Manfred. 1989. The semantics of gradation. In Manfred Bierwisch & Ewald Lang (eds.), *Dimensional Adjectives*, 71–261. Berlin: Springer Verlag.

Cresswell, Max J. 1977. The semantics of degree. In Barbara Partee (ed.), *Montague Grammar*, 261–292. New York: Academic Press.

Fox, Danny & Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy* 29(5). 537–586. doi:10.1007/s10988-006-9004-4.

Fults, Scott. 2006. *The structure of comparison: an investigation of gradable adjectives*: University of Maryland Ph.D. dissertation.

Heim, Irene. 1985. Notes on comparatives and related matters. Manuscript, University of Texas, Austin TX.

Heim, Irene. 2000. Degree operators and scope. In *Semantics and Linguistic Theory (SALT) X*, CLC Publications.

Kennedy, Christopher. 1999. *Projecting the adjective: the syntax and semantics of gradability and comparison.* Outstanding dissertations in linguistics. New York: Garland.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy* 30. 1–45. doi:10.1007/s10988-006-9008-0.

Kennedy, Christopher & Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language* 81. 1–37. doi:10.1353/lan.2005.0071.

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4(1). 1–45. doi:10.1007/BF00351812.

Klein, Ewan. 1982. The interpretation of adjectival comparatives. *Journal of Linguistics* 18. 113–136. doi:10.1017/S0022226700007271.

Klein, Ewan. 1991. Comparatives. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantics: An International Handbook of Contemporary Research*, 673–691. Berlin: de Gruyter.

Kranz, David H., R. Duncan Luce, Patrick Suppes & Amos Tversky. 1971. *Additive*

and Polynomial Representations, vol. I Foundations of Measurement. New York: Academic Press.

Lassiter, Dan. 2011. *Measurement and modality: the scalar basis of modal semantics*: New York University Ph.D. dissertation.

Link, Godehard. 1983. The logical analysis of plurals and mass terms. In Rainer Bäuerle, Christoph Schwarze & Arnim von Stechow (eds.), *Meaning, Use, and Interpretation of Language*, 302–323. Berlin: Mouton de Gruyter.

van Rooij, Robert. 2011. Measurement, and interadjective comparisons. *Journal of Semantics* 28(3). 335–358. doi:10.1093/jos/ffq018.

Sassoon, Galit. 2010. Measurement theory in linguistics. *Synthese* 174(1). 151–180. doi:10.1007/s11229-009-9687-5.

Schmidt, Lauren, Noah Goodman, David Barner & Joshua Tenenbaum. 2009. How tall is *tall*? Compositionality, statistics, and gradable adjectives. In *Annual Meeting of the Cognitive Science Society (COGSCI 2009)*, 3151–3156. Amsterdam.

Solt, Stephanie. 2011. Notes on the comparison class. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in Communication (ViC2009), Revised Selected Papers (LNAI 6517)*, 189–206. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-18446-8_11.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167. doi:10.3758/s13428-010-0039-7.

von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3. 1–77. doi:10.1093/jos/3.1-2.1.

Stevens, Stanley Smith. 1946. On the theory of scales of measurement. *Science* 103. 677–680. doi:10.1126/science.103.2684.677.

Stevens, Stanley Smith. 1975. *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*. New York: Wiley.

Toledo, Assaf & Galit W. Sassoon. 2011. Absolute vs. relative adjectives - variance within vs. between individuals. In Neil Ashton, Anca Chereches & David Lutz (eds.), *Semantics and Linguistic Theory (SALT) XXI*, 135–154.

Tribushinina, Elena. 2011. Once again on norms and comparison classes. *Linguistics* 49(3). 525–553. doi:10.1515/ling.2011.016.

Experimenting with degree

Stephanie Solt
FNWI, ILLC
University of Amsterdam
P.O. Box 94242
1090 GE Amsterdam
The Netherlands
stephanie.solt@gmail.com

Nicole Gotzner
Institut für Deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
Dorotheenstr. 24
10117 Berlin
Germany
nicole.gotzner@hu-berlin.de