# Color Adjectives, Standards and Thresholds: An Experimental Investigation*

Nat Hansen
Department of Philosophy, University of Reading

Emmanuel Chemla
Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)
Departement d'Etudes Cognitives (École Normale Supérieure – PSL* Research University)

January 7, 2015

**Abstract**

Are color adjectives ("red", "green", etc.) relative adjectives or absolute adjectives? In this paper we conduct two experiments, one based on entailment patterns and one based on presupposition accommodation, that investigate the typology of scalar adjectives. We find evidence that the "quantitative" reading of color adjectives is interpreted generally like paradigmatic minimum standard absolute adjectives ("spotted", e.g.), with the important exception that there is significant interpersonal variation in where on the scale the standard is located. We also find evidence that paradigmatic relative adjectives ("tall", "wide") have a lower "threshold" that must be crossed before we observe purely relative behavior (participants refuse to identify the taller of two objects as "the tall one" if they are both very short), and that there is variation in where this lower threshold is located. We propose a unified schematic structure for relative and absolute adjectives: adjectives behave like traditional relative adjectives for objects between a lower and an upper threshold on the scale, and they behave like absolute adjectives for values outside of this range. Traditional minimum standard and maximum standard absolute adjectives are obtained as the limit cases when these thresholds occupy extreme values. We discuss the relevance of these findings for debates about the nature and extent of semantic context sensitivity in which color adjectives have played a key role.

*Thanks to Shen-yi Liao, Eliot Michaelson, participants in the Semantic Content seminar at All Souls College, Oxford, the Philosophy Society at the University of Reading, the discussion group at 4 Les Gauguins, the members of CCCOM, and the Philosophie Expérimentale workshop at the Institut Jean-Nicod for very helpful comments and discussion. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC, and British Academy grant SQ120050, "Quantitative Methods in Experimental Philosophy of Language".

## 1 Introduction

Scalar adjectives like "long" and "expensive" have played a central role in recent debates in philosophy over how best to understand the interaction of context and linguistic meaning.[1] Scalar (or gradable) adjectives apply to objects that possess the relevant property to varying degrees: one lunch can be *longer* than another; one book can be *more expensive* than another. Semanticists have distinguished two types of scalar adjectives: in their bare positive (non-comparative) form, *relative* adjectives, like "long" and "expensive", are evaluated against context sensitive standards (of, e.g., length or price). In contrast, *absolute* adjectives like "full" or "spotted" have conventionally fixed reference points, and therefore display less context sensitivity than relative adjectives.

Given that color adjectives are a type of scalar adjective (an apple can be "redder than" another, or "very red", "perfectly red", "completely red", and so on), are they relative or absolute? If they are relative adjectives, then they will display at least as much semantic context sensitivity as adjectives like "long" and "expensive", which are standardly understood as applying to objects only relative to a contextually supplied standard. The degree of semantic context sensitivity displayed by color adjectives is of particular interest because color adjectives have played a central role in debates about the plausibility of *radical contextualism*, the view that all (or almost all) expressions in natural language are context sensitive, and that those effects of context on the semantic content of what is said can't be explained using the resources of truth conditional semantic theory (Travis, 2008a). If the context sensitivity of color adjectives can be explained using the kind of technique that already exists to handle the context sensitivity of relative scalar adjectives like "long" and "expensive", then color adjectives don't lend support to radical contextualism (Hansen, 2011; Kennedy and McNally, 2010; Szabó, 2001).

It has been recently argued that color adjectives are *absolute* adjectives, not relative adjectives. More specifically, Clapp (2012) has argued that color adjectives are *minimum-standard absolute adjectives*, which require an object to possess only a minimum degree of the relevant property for the adjective to apply to the object.[2] If color adjectives turn out to be absolute, rather than relative, then one of the resources for explaining the variability of color adjectives will not be available to the defenders of more moderate forms of contextualism.

In this paper, we evaluate whether color adjectives pattern with relative or minimum standard absolute adjectives by using two experimental methods of evaluating the semantic

[1] Stanley (2004) claims that contextualist analyses of "know" rest on the similarity of "know" with context sensitive scalar expressions like "tall", and he has argued against the similarity on the grounds that "know" does not take degree modifiers and does not have a comparative form. DeRose (2008) defends a contextualist analysis of "know" against critics by arguing that the standards governing scalar adjectives like "tall" are just as messy ("pluralistic") as those governing "know". Glanzberg (2007) proposes a contextualist analysis of predicates of personal taste ("tasty", "fun") on the grounds that they share grammatical features with scalar adjectives. Cappelen (2012) proposes a contextualist account of "intuitive" on the grounds that it is a scalar adjective.

[2] McNally 2011 offers a more subtle version of the view that color adjectives are absolute, which will be discussed below.

properties of scalar adjectives:

- Entailment tests (Burnett 2012, Kennedy 2007, Kennedy and McNally 2005, Toledo and Sassoon 2011)
- The presupposition accommodation task (Syrett et al., 2010)

We find that reactions to color adjectives in these tests display surprising patterns that diverge from existing armchair judgments. Once we distinguish *quantitative* and *qualitative* readings of color adjectives, reactions to the quantitative reading split roughly into three groups that differ in where the standard is located on the scale: one group responds as if the quantitative reading were *minimum standard absolute*, a second group responds as if a minimum-like absolute standard were located somewhere in the middle of the scale, and a third that responds as if the standard were located at the upper end of the scale. Also surprisingly, we find that participants respond to relative adjectives as if there were a *threshold* that an object has to cross before they accommodate the existence presupposition associated with the definite description in requests like "Please click on the red alien". And we find that participants differ significantly in terms of where they locate this threshold. In contrast to responses to minimum standard, relative, and the quantitative reading of color adjectives, the qualitative reading does not display any clear pattern of reactions—though it is possible to say that it does pattern significantly differently from both relative and minimum standard absolute adjectives, and that judgments of color quality vary across individuals. We discuss the relevance of these findings for larger debates about the nature and extent of semantic context sensitivity in which color adjectives have played a key role, and for the understanding of the typology of scalar adjectives in general.

### 1.1 Background on the semantics of scalar adjectives and the relative/absolute distinction

The primary tests used to distinguish scalar from non-scalar adjectives are whether the adjective can appear felicitously in comparative constructions (without coercion), and whether the adjective can combine with degree modifiers (e.g., "very"):

(1) The hardcover is more expensive than the paperback.

(2) The hardcover is very expensive.

(3) # The number seven is more prime than the number 5.

(4) # The number seven is very prime.

While non-scalar adjectives are associated with functions that map arguments to truth values, on a standard, "off the shelf" semantics for scalar adjectives, they are associated with functions from arguments to degrees on a scale (Bartsch and Vennemann, 1972; Kennedy, 2007; Syrett et al., 2010):[3]

[3] See Glanzberg (2007) for the "off the shelf" description.

(5) $[\![\textbf{prime}]\!]_{<e,t>} = \lambda$x . prime(x)

(6) $[\![\textbf{expensive}]\!]_{<e,d>} = \lambda$x . expensive(x)

Turning a scalar adjective plus argument into something that is truth-evaluable requires some kind of *comparison*. In a comparative construction, the comparison is explicit: "The hardcover is more expensive than the paperback" is true just in case the hardcover is mapped to a greater degree on the scale of *cost* than the paperback:

(7) $[\![\textbf{more G than}]\!]_{<<e,d>,<e,<e,t>>>} = \lambda$G$\lambda$y$\lambda$x . G(x)$\succ$ G(y)

(8) $[\![\textbf{more expensive than}]\!]_{<<e,<e,t>>} = \lambda$y$\lambda$x . expensive(x) $\succ$ expensive(y)

(9) $[\![\textbf{The hardcover is more expensive than the paperback}]\!]_t$ = expensive(the hardcover) $\succ$ expensive(the paperback)

When scalar adjectives occur without explicit comparative morphology, as in (10), a comparison is still involved, but it is implicit:

(10) The hardcover is expensive.

One way of allowing for the implicit comparison is to claim that when scalar adjectives appear without explicit comparative morphology, the adjective is accompanied by an unpronounced ("null") morpheme that supplies the relevant comparison. So, for the purposes of semantic interpretation, a "bare positive" construction like (10) is actually understood as (11):

(11) The hardcover is *pos* expensive.

*pos* supplies the scalar adjective it combines with with a context-sensitive function *standard*, which "chooses a standard of comparison in such a way as to ensure that the objects that the positive form is true of 'stand out' in the context of utterance, relative to the kind of measurement that the adjective encodes" (Kennedy, 2007, p. 17):

(12) $[\![\textbf{pos}]\!]_{<<e,d>,<e,t>>} = \lambda$G$\lambda$x . G(x) $\succeq$standard(G)

(13) $[\![\textbf{pos expensive}]\!]_{<e,t>} = \lambda$x . expensive(x) $\succeq$standard(expensive)

(14) $[\![\textbf{The hardcover is pos expensive}]\!]_t$ = expensive(The hardcover) $\succeq$standard(expensive)

So (14) is true just in case the hardcover is mapped to a degree of cost that "stands out" relative to cost in the context of utterance.

What is it for an object to "stand out" in terms of a kind of measurement? In the case of "expensive", what may stand out in terms of its cost in one context may not in another. In a context where the cost of different versions of some particular book is being assessed, then the hardcover stands out in terms of its cost. But in a different context (such as a discussion of what gift to get someone, for example) where the cost of the hardcover is being compared with the cost of a bottle of wine, then the hardcover might not "stand out" in terms of its cost.

Those adjectives for which it can vary across contexts whether an object counts as "standing out" in terms of the kind of measurement the adjective encodes are *relative* adjectives. The observation of the behavior of relative adjectives dates at least to 1632, in Galileo's *Dialogue Concerning the Two Chief World Systems*:

> I say that these terms 'large,' 'small', 'immense,' 'minute,' etc. are not absolute, but relative; the same thing in comparison with various others may be called at one time 'immense' and at another 'Imperceptible,' let alone 'small.'

More recently it has been argued that there is another category of scalar adjectives—*absolute* adjectives—that, unlike relative adjectives, don't display contextual variability in standards (Unger, 1975; Yoon, 1996; Rotstein and Winter, 2004; Kennedy and McNally, 2005; Kennedy, 2007; Syrett et al., 2010). Absolute adjectives have conventionally fixed standards, and have themselves been divided into two categories: *maximum standard* (or *total*) absolute adjectives, and *minimum standard* (or *partial*) absolute adjectives. Maximum standard absolute adjectives (e.g., "pure", "empty", "full", "flat") are associated with a standard fixed by the maximum degree on the scale associated with the adjective.[4] Minimum standard absolute adjectives (e.g., "impure", "visible", "spotted"), are associated with a standard fixed by the minimum degree on the scale associated with the adjective.

The different ways that the standard values of absolute and relative adjectives are determined is built into the lexical meaning of each adjective, and when combined with "pos", they generate different truth conditions, as follows (see Kennedy 2007, p. 26):

(15) pos adjective$_{min}$ (x) $\succ$ minimum degree on the scale associated with the adjective

(16) pos adjective$_{max}$ (x) = maximum degree on the scale associated with the adjective

(17) pos adjective$_{rel}$ (x) $\succeq$ contextually determined standard degree on the scale associated with the adjective

When a scalar adjective combines with "pos", whether the adjective is relative or minimum or maximum standard absolute is part of the input to the context sensitive "standard" function that is part of the meaning of "pos", which determines the adjective's standard value. With minimal and maximal standard absolute adjectives, the standard value is determined by the lexical meaning of the adjective alone (and remains fixed), while the standard value for relative adjectives can vary across contexts.

### 1.2 Hybrid standards

The standard, "off the shelf" semantics for scalar adjectives described in the previous section provides a neat taxonomy of different standards: *minimum standard absolute*, *maximum standard absolute*, and *relative*. There is, however, a more abstract way of thinking about standards from which the off the shelf picture can be derived, but which also allows

[4]For a criticism of the standard way of drawing the relative-absolute distinction, and an alternative proposal, see Toledo and Sassoon (2011).

theoretical space for a wider range of types of standard, and which makes some distinct empirical predictions from the off the shelf picture.

On the more abstract picture of standards for scalar adjectives, all adjectives share the same general type of standard, which is composed of three elements: a *lower threshold*, an *upper threshold*, and a *standard* (see Figure 1).

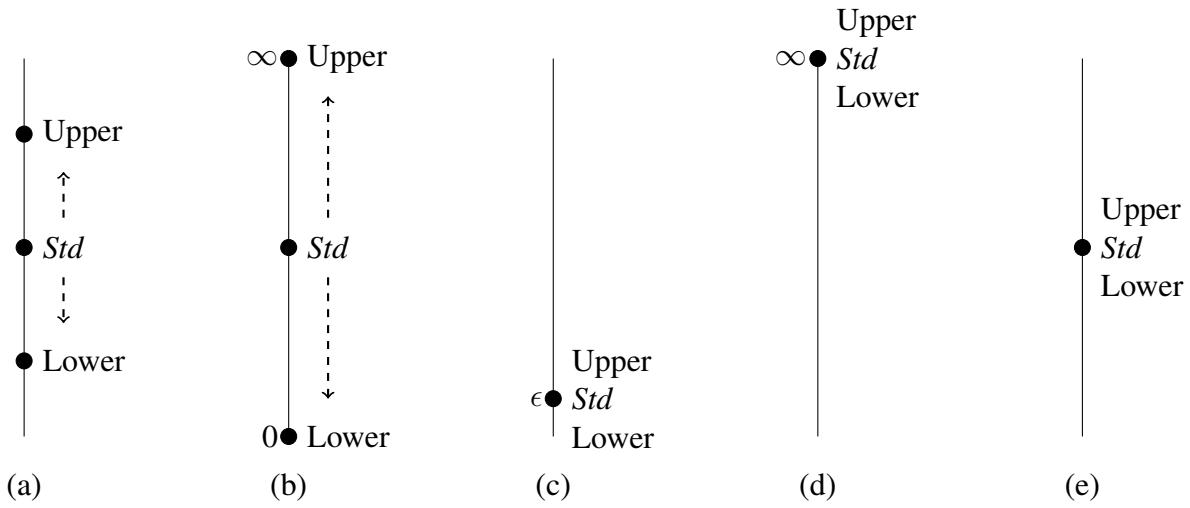



Figure 1: A typology of adjectives based on a lower threshold (Lower) and an upper threshold (Upper), which together delimit the area of the scale where an adjective is relative, with a contextually variable standard in between (*Std*). The general, hybrid structure of adjectives is given in (a). Traditional relative adjectives are obtained when Upper and Lower are at the extremes of the scale (b). If Upper and Lower collapse (in which case the intermediate standard is also forced in this position) there is no area where the adjective behaves 'relatively', and we obtain minimum (c), maximum (d) or intermediate (e) absolute standard adjectives (see McNally 2011 for the latter).

To motivate the more abstract structure of standards, consider a situation where there are two tiny toy soldiers, one of which is noticeably taller than the other. If you ask me to hand you *the tall soldier*, I might reasonably object, on the grounds that neither is sufficiently tall to count as *the tall soldier*. This behavior runs counter to what is predicted by the standard picture, which holds that "tall" has a relative standard that can be shifted by the process of presupposition accommodation to pick out the *taller* of the two soldiers, however tall or short they may be (this prediction and its justification will be discussed in §4). In this situation, neither toy soldier meets or exceeds the lower threshold.

A parallel situation might obtain at the upper end of the scale as well. Imagine a situation in which there are two giant sequoia trees, towering over everything else around, but one of the trees is noticeably taller than the other. In such a situation, if you asked me to *take a picture of the tall tree*, I might reasonably call for clarification or object to your request on the grounds that both sequoias are tall. In this situation, both trees meet or exceed the upper threshold.

These two imagined situations involve linguistic behavior that is characteristic of maximum standard absolute adjectives (when the objects fall below the lower threshold), or minimum standard absolute adjectives (when the objects rise above the upper threshold). But if the relevant objects are associated with a degree on the relevant scale that is *between* the lower and upper thresholds, then the adjective will behave like a standard relative adjective. So, for example, if we're deciding which of two people to guard in a soccer game, one of which is taller than the other, but neither one of which is extremely short or extremely tall, you can tell me to guard the taller of the two by saying *guard the tall one*. Adjectives that display this pattern of behavior, characterized by features of maximum standard, minimum standard, and relative adjectives can be said to have *hybrid* standards.

It is possible to derive all of the traditional types of standard from this more abstract structure of standards. The behavior of traditional relative standards would result from setting the lower threshold at the minimum degree of the relevant scale, and upper threshold at infinity (see Figure 1b). A traditional minimum absolute standard is equivalent to collapsing the lower and upper thresholds at the minimal (but non-zero) degree on the scale (see Figure 1c). And a traditional maximum absolute standard is equivalent to collapsing the lower and upper thresholds at the maximum degree on the scale (see Figure 1d).

While the more abstract picture of the structure of standards allows the derivation of the traditional picture, it also allows for the possibility of a variety of hybrid standards. For example, McNally (2011) discusses the possibility of absolute standards that aren't located at either the minimal or maximal degrees of a scale. Notably, she suggests that color adjectives (on their quantitative reading, which will be discussed in the next section) are associated with such an absolute standard: an object counts as, e.g., red just in case red is the *predominating* color of the object. This standard is not contextually variable in the way that the standards of relative adjectives are, and it's located somewhere in the middle of the scale. On the abstract picture of standards, McNally's middle-of-the-scale-absolute standard would in effect be one where the two thresholds and the standard are collapsed in the middle of the scale, as in Figure 1e.

Consistent with (but not entailed by) the more abstract picture of standards is the strong view that every adjective is hybrid—that is, there is always some gap between the lower and upper thresholds in which the adjective will behave like a relative adjective. (The gap between the two thresholds might be small, which would require subtle tests to uncover.) A weaker view would allow for the existence of the traditional absolute standards as well as intermediate absolute standards like the one proposed by McNally (which result from the collapsing of the lower and upper thresholds), but also for the existence of hybrid standards.

This unified, more abstract picture of standards can recede into the background until we get to the results of our second experiment, which lends some support to the existence of hybrid standards.

### 1.3 Color adjectives: Background

Radical contextualists have argued that color adjectives are highly context-sensitive expressions the variation of which can't be adequately explained using the resources of traditional

compositional truth conditional semantic theory (see Travis 2008b, Kennedy and McNally 2010 and Hansen 2011 for discussion).[5] Defenders of a variety of more moderate semantic theories have used non-radical resources, like indexicality and ambiguity, to explain the contextual variability of the content of color adjectives (Hansen, 2011; Kennedy and McNally, 2010; Rothschild and Segal, 2009; Szabó, 2001; Vicente, 2015).

Kennedy and McNally (2010) relate the debate over the radical context sensitivity of color adjectives to semantic theories of scalar adjectives. They distinguish two types of ambiguity that characterize adjectives in their positive form: first, they argue that there is a non-scalar use of color adjectives, in which there is an all-or-nothing correlation between the color and some contextually relevant property that the color is an indicator of. For example, blue flags are used to indicate beaches that have "met an international standard for a clean and healthy beach".[6] One might refer to such a beach by saying (18), thereby expressing the proposition that the beach has met a particular international standard for cleanliness:

(18) This beach is blue.

Kennedy and McNally argue that this interpretation of color adjectives can't felicitously appear in comparative constructions or take degree modifiers, because the property being expressed isn't associated with a scale.

The second type of ambiguity that Kennedy and McNally observed is a characteristic of scalar color adjectives, and emerges when one considers the way color adjectives interact with degree modifiers. It is possible to tease apart two different scales associated with color adjectives, like "green", when one considers that an object can be *completely* green without being *perfectly* green, and vice versa. Both "completely" and "perfectly" pick out maximal degrees on the scale of the adjective they modify, but they appear to be modifying different scales, as (19–20) demonstrate:

(19) The leaf is too yellowish to be perfectly green, but it is completely green.

(20) The leaf is only 90% green, but it is perfectly green.

Sentence (19) shows that an object can have less than a maximal degree on the scale of *qualitative* greenness even if the object is *completely* quantitatively green. And (20) shows that something can have a maximal value on the scale of qualitative greenness without having a maximal degree on the scale of quantitative greenness.

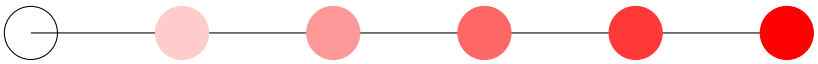
There is therefore reason to think that color adjectives have two distinct scalar readings: a *quantitative* reading (Figure 2), which is associated with a scale of how much of an object is a particular color, and a *qualitative* reading (Figure 3) associated with a scale measuring "how closely an object's color approximates or diverges from a 'center' or prototype" (Kennedy and McNally, 2010, p. 91).

[5] Hansen and Chemla (2013) confirmed the armchair judgments of radical contextualists by demonstrating in experimental conditions that there is an effect of changing contexts on truth value judgments about sentences containing color adjectives ("black", "beige", "green", "red"). See Davies (2015) for critical discussion of this way of understanding radical contextualism.

[6] `http://www.blueflag.org/`.

Figure 2: Quantitative scale of redness


Figure 3: Qualitative scale of redness




While there is much to be said about the adequacy of Kennedy and McNally's account of color adjectives as a response to radical contextualism (see Clapp 2012, Davies 2015 and Hansen 2011 for discussion), our focus in this paper will be on working out whether these two scalar readings of color adjectives have absolute or relative standard values (or some other kind of standard), and what that tells us about the context sensitivity of color adjectives and about how to understand the nature of standards as they apply to scalar adjectives.

## 2 Are color adjectives more like "tall" or "spotted"?

Clapp (2012) defends a type of radical contextualism about the truth conditional variation of sentences containing color adjectives. An interesting part of his defense of that view is an argument that color adjectives are *absolute* scalar adjectives, and that they therefore do not display the kind of semantic context sensitivity characteristic of relative scalar adjectives like "tall" and "expensive". That is supposed to cut against accounts of color adjectives, like the one given in Hansen (2011), which seem to assume that color adjectives are semantically context sensitive relative adjectives.[7] As described above, relative scalar adjectives, when they appear in positive form, have context dependent standards, whereas absolute adjectives have standards that are set by the lexical meaning of the adjective (either at the maximum degree on the associated scale, or just above the minimum degree). Clapp suggests that color adjectives are minimum-standard absolute adjectives, like "spotted", where an object counts as having the relevant property just in case it is projected to a degree on the scale associated with the adjective that exceeds the minimum degree of the scale (p. 97). If he's right, then color adjectives would not display the form of semantic context sensitivity

[7] I say "seem to assume", because Hansen (2011) does not take an explicit stand on whether color adjectives are relative or absolute. Kennedy and McNally (2010, p. 92) say that the qualitative reading of color adjectives can vary with differences in comparison class, a characteristic of relative and not absolute adjectives. But regarding the quantitative reading, they say that "judgments are surprisingly consistent as to how much of an object must manifest the color in order for the term to be applicable. Our preliminary observations suggest that in order for a color term to apply to an object on the scalar quantity reading, the color in question must perceptually predominate". The claim that there is a fixed standard for the quantitative reading makes it sound like an absolute adjective, albeit one where the standard isn't anchored by the minimum or maximum degree on the scale.

specific to relative adjectives.[8]

Clapp makes two arguments in support of the claim that color adjectives have absolute standards. His first argument relies on distributional evidence that is supposed to show that the structure of the scale associated with color adjectives has maximal and minimal degrees, and on the controversial "Interpretive Economy Principle" (Kennedy, 2007) that associates absolute standards with closed scales. His second argument relies on the "presupposition assessment task" developed in Syrett et al. (2010), which is used to show that while speakers accommodate the existence and uniqueness presuppositions of definite descriptions when combined with relative adjectives (as in "Please hand me the long stick"), speakers refused to accommodate when definite descriptions are combined with absolute adjectives (as in "Please hand me the spotted disk"). Clapp claims, based on his own armchair judgments, that color adjectives behave like absolute adjectives in the presupposition assessment task. There is also another type of test, involving entailment patterns, that can be applied to color adjectives to determine whether they pattern with relative or minimum standard absolute adjectives. We will briefly discuss Clapp's first argument before going on to discuss experiments involving entailment patterns and the presupposition assessment task.

**Distributional data and the interpretive economy principle**

There is intuitive support for associating color adjectives with scales that are "closed", that is, that have both minimal and maximal degrees. Objects that are not at all red (if they're completely green, or yellow, or blue, achromatic, etc.) would not be associated with a degree on the scale of redness. So it seems that there should be a minimum degree on both the quantitative and qualitative scales of redness. A saturated, bright, unique red would occupy the maximum degree on the qualitative scale of redness, as illustrated by the rightmost circle in Figure 3. And the maximum degree on the *quantitative* scale of redness would be occupied by an object that is completely red, as illustrated by the rightmost circle in Figure 2.

The way that degree modifiers combine with color terms also supports the claim that they have maximal and minimal degrees (Clapp 2012, pp. 95–96). The modifier "slightly" tends to be more acceptable with minimum standard absolute adjectives than with relative adjectives (Rotstein and Winter, 2004), and "slightly" seems to combine felicitously with color adjectives:

(21) His face is slightly red.[9]

[8]Showing that color adjectives are absolute would not show that color adjectives are not semantically context sensitive. As indicated in Kennedy and McNally (2010) and Hansen (2011), it might be the case that what *scale* is associated with a particular color adjective is itself context sensitive, even if the *standard* is conventionally fixed at some degree on the scale.

[9]Burnett (2012, p. 7) (citing Solt 2011) observes that "slightly" combines easily with both minimum standard absolute adjectives and relative adjectives (where it is naturally interpreted as "slightly too"):

**Minimum standard absolute adjectives**

Secondly, it seems that proportional modifiers like "50%", "mostly", and "two thirds" can combine felicitously with color adjectives, and proportional modifiers require the adjectives that they modify to have both minimal and maximal degrees (otherwise there would be no way to calculate a midpoint on the scale, for example).

(22) The camouflage is half green, half brown.

Finally, as discussed in §1.3 above, color adjectives can be combined with maximal degree modifiers like "completely" and "perfectly", which is evidence that they are associated with scales that have maximal degrees.

Assuming for the time being that Clapp is right about this distributional evidence and the structure of the scales associated with color adjectives, what does it tell us about the status of color adjectives as relative or absolute? Kennedy and McNally (2005, p. 361) conditionally "assume that interpretations that minimize context-dependence are in general preferred", and observe that "the endpoints of a totally or partially closed scale provide a fixed value as a potential standard", thereby providing the relevant context-independent standard. Kennedy goes on to explain the connection between whether a scale has endpoints and its status as relative or absolute in terms of his *Interpretive Economy Principle*:

> *Interpretive Economy*
>
> Maximize the contribution of the conventional meanings of the elements of a sentence to the computation of its truth conditions ...
>
> The effect of *Interpretive Economy* on the positive form is to ensure that closed scale adjectives are absolute... A context-dependent, relative standard of comparison is also in principle an option, but since an adjective's scale structure is part of its conventional meaning, *Interpretive Economy* dictates that the absolute truth conditions are the ones that should surface (Kennedy, 2007, pp. 36–37)

Combining the distributional evidence with the Interpretive Economy Principle yields the following argument:

1. Color adjectives are associated with closed scales (scales with minimal and maximal degrees).

---

(1) The floor is slightly dirty.

(2) The window is slightly open.

**Relative adjectives**

(3) My hair is slightly long.

(4) Dinner was slightly expensive.

When color adjectives combine with "slightly", they don't seem to receive the "slightly too" interpretation:

(5) The sunset is slightly pink.

(6) My hair is slightly gray.

2. The Interpretive Economy Principle holds that adjectives with closed scales should be absolute.

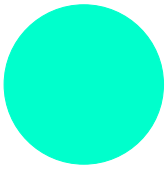
3. So color adjectives are absolute.

There is an important problem with the second piece of distributional evidence cited by Clapp, that color adjectives can combine with proportional modifiers like "half" or "mostly". He is right that the *quantitative* reading of color adjectives can combine with proportional modifiers:

(23) The flag of the PRC is mostly red.

But he doesn't distinguish between the quantitative and qualitative readings of color adjectives, and so misses the fact that the qualitative reading of color adjectives cannot combine with proportional modifiers.[10] For example, it isn't possible to describe the color quality of a bluish green circle (see Figure 4) by saying that it is "half green, half blue".[11] The distinction between the two readings of scalar color adjectives will be important when we discuss our two experiments that aim to determine what kind of standard is associated with color adjectives.

Figure 4: #The circle is half green, half blue.



Even if we grant that Clapp's remaining distributional evidence is sufficient to support the claim that both readings of color adjectives are associated with closed scales, there is a good reason to resist the conclusion of the argument (see McNally 2011, p. 156). The main problem with Clapp's first argument is that there are reasons to doubt that all adjectives associated with closed scales have absolute standards. Kennedy himself gives the example of "bald", which has a maximal degree ("completely bald"), but which varies its standard of comparison (where the cut-off for baldness is) in different contexts in the way relative adjectives do (Kennedy, 2007, p. 35 n. 30). McNally (2011) argues that "familiar" has a maximal degree ("completely familiar"), but has a standard that can shift in different contexts. Lassiter (2010) argues that the scalar epistemic modals "probable" and "likely" are associated with fully closed scales, but have relative standards. "Reliable"/"unreliable"

[10]The fact that the qualitative reading does not combine with proportional modifiers is pointed out by Kennedy and McNally (2010, p. 92).

[11]Kennedy and McNally made this point using a turquoise Honda in a talk given at the International Conference on Adjectives at the Université de Lille, 2007. If one is working with a technical measure of color quality, then one can say of a shade, e.g., that it is 70% red, 30% yellow, in effect turning color quality into a form of color quantity.

also appear to fit into this category of closed scale adjectives that have relative standards—something can be *completely reliable* or *completely unreliable*, but the standard for what counts as reliable isn't fixed by the maximal or minimal degree on the scale of reliability. So even if it is generally true that adjectives with closed scales have absolute standards, and that is a correlation that requires some explanation, there is room to doubt that the fact that color adjectives are associated with closed scales means that they have contextually invariable absolute standards.

## 3 Experiment #1: Evidence from entailment patterns

One standard way of distinguishing relative from absolute adjectives is in terms of the entailments they can figure in. Minimum standard absolute adjectives require objects to possess only a minimal degree of the relevant property to count as having that property. So, for example, an object satisfies "is spotted" just in case it is spotted *to some degree*. But relative adjectives are different: how much of the relevant property an object needs to possess before it counts as having the relevant property can vary across contexts, and an object must have more than a minimal degree of the relevant property to count as satisfying the adjective. Being tall, for example, obviously requires more than just having more than a minimal degree of height.

Accordingly, minimum standard absolute adjectives, but *not* relative adjectives, support the following entailments (Burnett, 2012; Kennedy and McNally, 2005; Kennedy, 2007; Toledo and Sassoon, 2011):

(24) **Entailment pattern #1**
X is more $\text{Adj}_{min}$ than y $\Rightarrow$ X is Adj.

That is, if x has more of property Adj than y, then x has a non-minimal degree of property Adj. On the standard picture of minimum standard absolute adjectives, x only requires a minimal degree of Adj-ness to satisfy the adjective, so x is Adj.[12]

(25) **Entailment pattern #2**
X is not $\text{Adj}_{min}$ $\Rightarrow$ x has a zero degree of Adj-ness.

That is, if all it takes for x to have property Adj is to have a minimal degree of Adj-ness, then not having that property means it has no degrees of Adj-ness. But the entailment clearly doesn't hold for relative adjectives: "x is not tall" does not entail that x has zero degrees of height!

[12] Burnett (2012, pp. 8–9) questions whether minimum standard absolute adjectives in the comparative always have this entailment. She cites the example of "dangerous", which doesn't have the entailment:

(1) Driving from Ottawa to Toronto is more dangerous than flying from Ottawa to Toronto #Driving from Ottawa to Toronto is dangerous.

But this seems like evidence that "dangerous" is *not* a minimum standard absolute adjective, but a relative adjective. Similarly, "x is not dangerous" doesn't entail "x has zero degrees of dangerousness", which is further evidence that "dangerous" is not minimum standard absolute.

Do color adjectives pattern with minimum standard absolute adjectives or like relative adjectives with regard to these entailments? We conducted an experiment that aimed to (1) confirm existing armchair judgments about the different entailment patterns that relative and minimum standard absolute adjectives are supposed to figure in, and (2) determine whether or not color adjectives display similar patterns. Note that both armchair judgments and the responses of participants in formal experiments are not direct evidence of entailments, but of *inference* patterns (that is, how speakers reason with language, rather than the logical properties of the language itself). But on the assumption that knowledge of the language (which includes entailment relations) guides speakers' linguistic judgments, then inference patterns are evidence of entailment patterns, unless there is reason to think some other factor is influencing inference patterns.

### 3.1 Materials, design and task

#### 3.1.1 Adjectives

We tested six adjectives of each of three types: minimum standard absolute, relative, and color (see Table 1).

| | | | | | | |
|---|---|---|---|---|---|---|
| *minimum standard* | bumpy | dirty | sick | spotted | visible | wet |
| *relative* | big | heavy | long | old | tall | wide |
| *color* | blue | brown | green | pink | red | yellow |

Table 1: Target adjectives in the entailment experiment

#### 3.1.2 Three inferential tasks

To test the entailment patterns that different types of adjectives figure in, we used three different inferential tasks for each of the two entailment patterns discussed above in (24) and (25).

- The downward arrow task ("↓") is intended to elicit more or less direct judgments about entailment. After a brief introduction to entailment, participants were asked to say whether a sentence following the downward arrow has to be true if the sentence preceding the arrow is true, as in (26a).

- The THEREFORE task is a linguistic translation of the "↓" test: participants were asked to say whether a sentence of the form "$p$ therefore $q$" makes sense, with $p$ and $q$ the appropriate premise and conclusion as in (26b).

- The BUT task was an anti-inference test, in which participants were asked to say whether a sentence of the form "$p$, but not $q$" makes sense, see (26c); negative responses here indicates entailment from $p$ to $q$.

(26) The inferential tasks, illustrated with the adjective 'tall' and the first entailment pattern (see (24)).

a. Downward arrow task "↓":

"X1 is taller than Y2."

↓

"X1 is tall."

b. **THEREFORE** task:
"X1 is taller than Y2, therefore X1 is tall."

c. **BUT** task:
"X1 is taller than Y2, but X1 isn't tall."

For all tasks, participants could indicate their response by clicking on either "yes" or "no".

### 3.1.3 Order of presentation

Each participant was presented with all possible combinations, for a total of 3 adjective types × 6 adjectives × 2 entailment patterns × 3 inferential tasks =108 test items.[13]

Because the "↓" inference test required different instructions from the BUT and THEREFORE tests, we divided the experiment into two "blocks", one containing the "↓" conditions and one containing the BUT and THEREFORE conditions. Test items within each block were randomized, following irrelevant training items (which were included to let participants get used to the display), and participants were randomly assigned to either a "↓"–first or "↓"-second ordering of the blocks. We observed no order effects of blocks.

### 3.1.4 Predictions

We were interested in how color adjectives would behave. Minimum standard adjectives should verify both entailment patterns, relative adjectives should not verify either of them, leading to the predictions in Table 2.

| | relative | minimum | color |
|---|---|---|---|
| "↓" | no | yes | ? |
| THEREFORE | no | yes | ? |
| BUT | yes | no | ? |

Table 2: Predictions for the inferential tasks

[13]Due to a coding error, the following conditions were not displayed: In the "↓" test, "dirty", "spotted", and "visible" were omitted from the "↓"-second order of the blocks.

### 3.2 Participants

41 participants were recruited over Amazon Mechanical Turk, and paid $0.80 each. One participant did not report to be a native English speaker and was therefore excluded from the analyses. Ages ranged from 21 to 66. 19 participants were female, and 22 male.

### 3.3 Results of the entailment pattern experiment

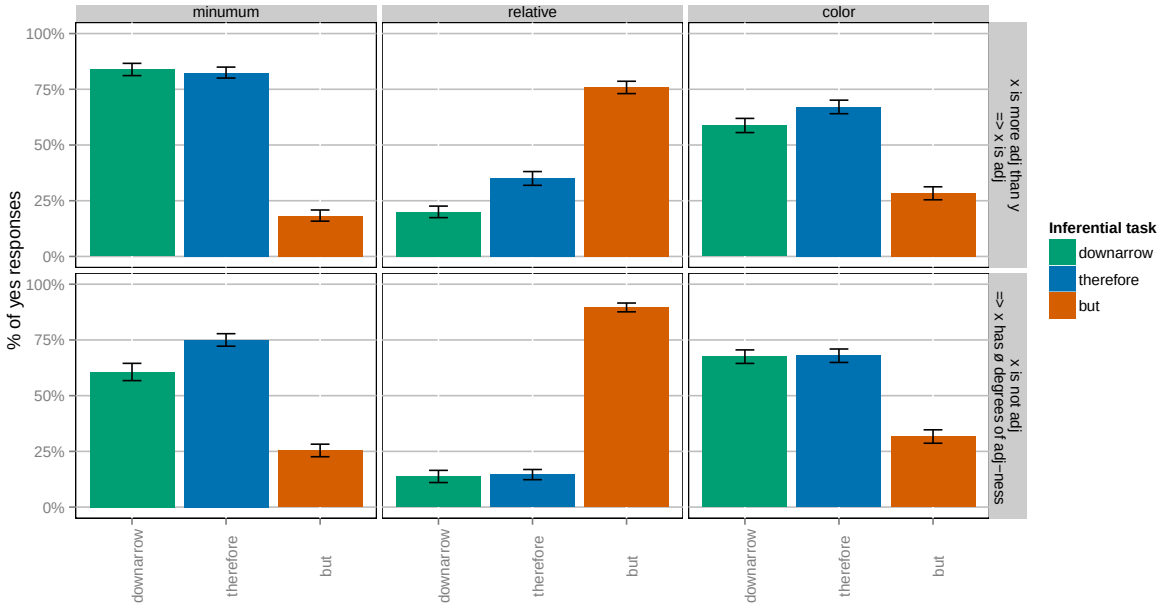



Figure 5: Percentages of "Yes" responses to the entailment pattern experiment

Four clear results are visible in Figure 5, which reports mean proportion of "YES" responses to the test items. First, as expected, responses to the BUT inference task are the mirror image of the responses in the THEREFORE/"↓" tasks. Second, all tests show that the two entailment patterns are generally accepted with minimum standard adjectives but not with relative adjectives and there are clear differences between the two in all conditions (all $p$s below .0001).[14] Hence the entailment patterns do distinguish between those two types of adjectives, as predicted in the literature. Third, responses to color adjectives clearly display different patterns of responses than relative adjectives for both types of entailment patterns, in all three inference tests (all $p$s below .0001). Fourth, we didn't find evidence that responses to color adjectives differ from minimum standard absolute adjectives: among the six different comparisons the most favorable $p$-values (in terms of trying to establish a

[14]Throughout the paper, we report results of anovas comparing models with and without the relevant predictor (here, adjective type), using logit models with participant as a random factor with slope and intercept. We do not include random factors for items because we had very few repetitions (here, 6 per condition) and could therefore not draw meaningful statistical generalizations about items. In essence, we therefore report per subject analyses; we have visually checked that there was no obvious oddball in our set of items.

difference) are obtained for entailment pattern 2 (.042 for the "↓" and .11 for BUT), but this outcome would not pass correction for multiple comparison.[15]

### 3.4 Discussion

First, this experiment confirms that the two entailment patterns clearly distinguish between minimum standard absolute and relative adjectives. So it provides a non-armchair confirmation of this widely-cited diagnostic. Second, most interestingly for our purposes, responses to color adjectives are significantly different than responses to relative adjectives.

One important limitation of this experiment is that it does not disambiguate the quantitative and qualitative readings of color adjectives. Without a way of disambiguating the two readings, the fact that responses to color adjectives are significantly different than responses to relative adjectives might be due to the fact that, for example, the quantitative reading is minimum standard absolute while the qualitative reading is relative.

Our second experiment addresses this limitation by disambiguating the two readings of color adjectives and allowing for a finer-grained assessment of the standards involved in minimum standard absolute, relative, and color adjectives.

## 4 Experiment #2: The presupposition assessment task

Following a suggestion in Clapp (2012), we employ versions of the experiments described in Syrett et al. (2010) which support the distinction between absolute and relative adjectives to assess whether color adjectives behave like absolute adjectives or relative adjectives (or something in-between). The experiments in Syrett et al. (2010) involve what they call a *presupposition assessment task*:

> Consider a situation in which two individuals A and B are sitting across from each other at a table, there are two blue rods of unequal lengths on the table in front of B [see Figure 6a], and A's goal is to get B to pass over one of the rods. In such a context, A cannot felicitously use [(27)] to make this request, because the existence presupposition is not met: there is no object that satisfies the property *red rod* in the context... By the same token, A's utterance of [(28)]

[15] One initially puzzling result visible in Figure 5 is that responses to relative adjectives in the THEREFORE task of the first entailment pattern seem to differ from responses to relative adjectives in the "↓" task. This result does not pass a proper significance test though ($p = .094$, before correction for multiple comparisons). But even if it did, this could be explained as an artifact of the instructions in combination with the test material. The instructions asked participants to indicate whether they thought that target sentences "made sense" or not. The test items for relative adjectives in this condition had the following form:

(1) X is taller than y, therefore x is tall.

While it doesn't follow with necessity from "X is taller than y" that "X is tall", saying (1) does *make sense* (that is, it isn't contradictory in the way that saying "X is completely clean but is covered in disgusting filth" is contradictory, which was how we illustrated that something "doesn't make sense" in the experimental instructions).

would also be infelicitous, in this case because the uniqueness presupposition of the definite description *the blue rod* is not met: there are two objects in the context that satisfy the property *blue rod*. Speaker A can, however, felicitously use [(29)] to request the longer of the two rods.

(27) # Please give me the red rod.

(28) # Please give me the blue rod.

(29) Please give me the long rod (Syrett et al., 2010, p. 5).[16]

a) Please give me the long rod b) Please give me the full one c) Please give me the spotted one

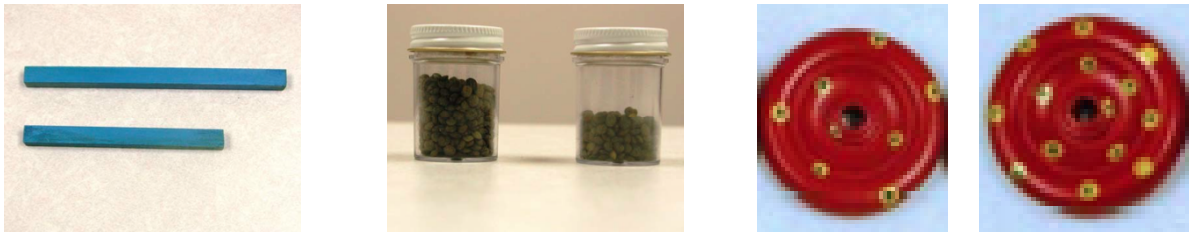


Figure 6: Examples from Syrett (2007, Appendix E)

Syrett et al. (2010) found that speakers were willing to comply with requests that involved accommodating the existence and uniqueness presuppositions of definite descriptions involving relative adjectives like "long", but would not do so for requests with definite descriptions involving absolute adjectives.

Consider the pair of jars in Figure 6b. If a competent speaker is asked by the experimenter to "Please give me the full one", which jar would be handed over? In a surprising confirmation of the distinction between absolute and relative adjectives, Syrett et al. (2010, p. 14) found that 88% of adult participants *rejected* the request to "Please give me the full one" in an experiment involving the two jars (where either handing neither or handing both counted as failure to accommodate). Only 12% responded to the request by handing over the *fuller* of the two jars. In contrast, 100% of adult participants responded to the request for the long rod by handing over the longer rod.[17]

In addition to the evidence of a refusal to accommodate the *existence* presupposition of the definite description in the request involving the maximum standard absolute adjective "full", there is also evidence of a failure to accommodate the *uniqueness* presupposition in a request involving the minimum standard absolute adjective "spotted". In Syrett et al.

[16]The numbering of examples has been brought into alignment with those in the current paper.

[17]In a second experiment designed to evaluate a possible order of presentation bias in the first experiment, Syrett et al. (2010, p. 17 n. 11) report that 70% of adults rejected the request for "the full one", down from 88% in the first experiment, but that "Adults who [complied by giving] the fuller of the two containers noted at the end of the experimental session without any prompting that they realized their mistake later in the experiment and wished to make clear to the experimenter that they knew what *full* means"!

(2010), 96% of adults rejected the request for "the spotted one" when both disks had some spots on them (as in Figure 6c).

Given the difference in meaning between relative and absolute adjectives discussed above, these different types of responses can be explained in terms of whether or not the adjective in the definite description has a standard value that is contextually variable (as with relative adjectives), and therefore capable of being shifted through the process of accommodation, or whether the standard value is fixed by the meaning of the adjective and therefore resistant to accomodation.

Clapp (p. 97) asks how people would respond to similar requests involving color adjectives. He says:

> . . . intuition suggests that competent interpreters are unable to accommodate definite descriptions involving color adjectives just as they are unable to accommodate absolute adjectives such as "spotted". That is, in a context containing two red objects, though one noticeably more red than the other, competent interpreters would reject a request made using
>
> (30) Please hand me the red one.
>
> as infelicitous.

According to Clapp, that would be evidence that color adjectives are minimum standard absolute adjectives, which means that an object would count as having the particular color expressed by the adjective just in case it exceeds the minimal semantically encoded standard, which does not vary across contexts, and thus cannot be adjusted through accommodation.

We found Clapp's intuition about the behavior of color adjectives surprising, and it did not accord with at least one of our (NH) own armchair judgments about how we would respond to the request to hand someone "the red one" when the two objects were both red, but clearly differed in terms of the *quality* of their redness. But one of the authors (again, NH) did share Clapp's judgment when it was the *quantitative* reading that was at issue. But these armchair judgments were nowhere near certain enough to convince us that we had the correct account of the standards associated with the two readings of color adjectives, and they also conflicted with the judgments about the quantitative reading of color adjectives given in Kennedy and McNally (2010) and McNally (2011), that the quantitative reading does behave like an absolute adjective, but only once the relevant color *predominates*. Adequately assessing these different predictions requires getting out of the armchair and conducting a formal experiment of how competent speakers respond to color adjectives in the presupposition assessment task.

### 4.1 Materials, Design, and Task

The aim of our second experiment is to evaluate whether subjects' responses to the qualitative and quantitative readings of color adjectives in the presupposition assessment task pattern with relative or minimum standard absolute adjectives, and whether they uncover

any evidence of the existence of thresholds for accommodation. As in Syrett et al.'s version, our task involves presenting subjects with two objects and asking them to select one of the objects or indicate their refusal to perform the task. In order to limit prototype effects and make the assignment of arbitrary colors to objects somewhat plausible, we asked subjects to respond to pictures of aliens and two refusal options, one indicating failure of the existence presupposition of the definite description ("Neither is!") and the other indicating failure of the uniqueness presupposition of the definite description ("Both are!") (see Figure 7).

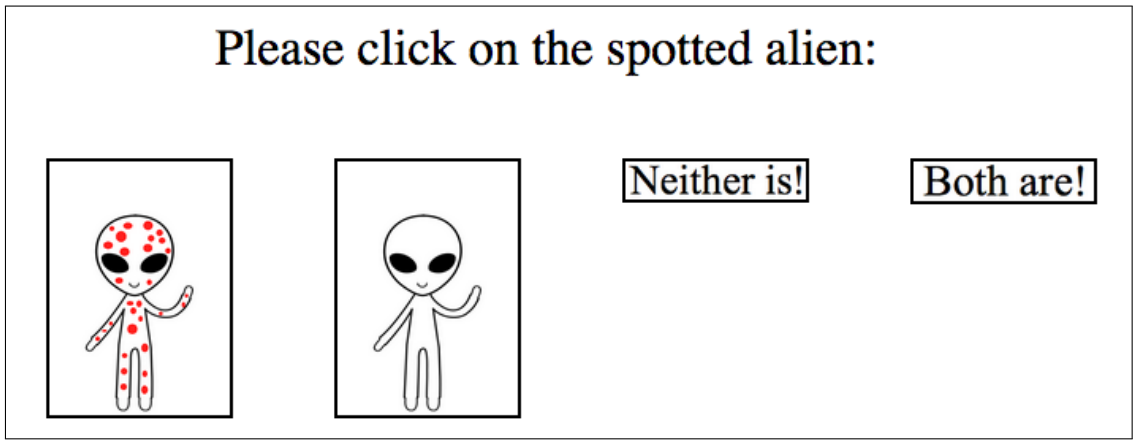



Figure 7: Alien selection task

#### 4.1.1 Adjectives and adjective types

The experiment involved four adjective types (relative, minimum standard, color quantity and color quality). The relative and minimum standard adjective types had two target words each ("tall" and "wide" for relative and "spotted" and "dirty" for minimum standard). The target words were selected because they were suitable for visual presentation (it would have been harder to test "expensive" or "wet", for example). The color quantity and color quantity adjective types each had four target words ("blue", "green", "red", and "yellow").

#### 4.1.2 Raw material: aliens with different degrees of adj-ness

The conditions were composed of two aliens. Individual aliens were thus created, satisfying the adjectives to different degrees, as illustrated in Figure 8, and as described below:

- Each adjective in the experiment is associated with a scale. A *maximal* condition was identified for each adjective. So, for the color quantity "red", the maximal alien was a completely red alien. For the color quality "red", the maximal alien was (what the experimenters judged to be) a focal ("best") example of redness. Relative and minimum standard adjectives do not have a genuine maximal degree, but the boxes surrounding the aliens (an artifact of the Ibex experimental program we used to create the experiments) provided a *de facto* maximum degree for both height and width: the

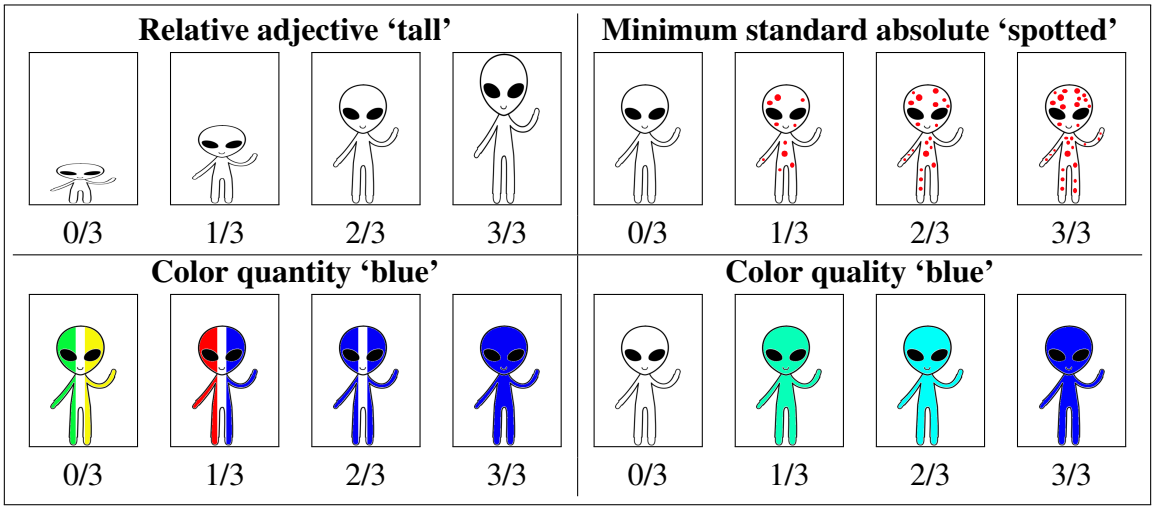



Figure 8: Examples of aliens used to construct the different conditions for adjectives of each of the four relevant types

maximally tall alien was as tall as the box, and the maximally wide alien was as wide as the box. For the minimum standard adjectives, we picked an arbitrary maximum degree (30 spots, in the case of "spotted", e.g.).

- Once the maximal degree for each target word was determined, we then created two less-than-maximal degree versions for each target word, in a "2/3" version and a "1/3" version. In the color quantity case, generating these aliens involved literally dividing the alien into (roughly) thirds and giving 2/3 or 1/3 of it the relevant color. For color quality, choosing the 2/3 and 1/3 versions of the aliens for each color term was more subjective. We aimed, in the 2/3 condition, to find a less paradigmatic example of the color that subjects would still be able to clearly categorize as an example of the relevant color (that is, not a borderline case). It turned out that finding appropriate 2/3 examples of "red" and "green" required varying the brightness dimension, yielding a darker red and a darker green than in the maximal, 3/3 condition, while appropriate 2/3 examples of "yellow" and "blue" involved varying the hue dimension, yielding a mustardy yellow and a Cambridge (as opposed to Oxford) blue, respectively. The 1/3 alien for color quality is intended to be a borderline example of the relevant color. For the relative and minimum standard adjectives, the 2/3 and 1/3 stimuli were generated in a straightforwardly proportional way: the 2/3 "tall" alien was 2/3 the height of the box, the 1/3 "tall" alien was 1/3 the height of the box (*mutatis mutandis* for "wide"), the 2/3 condition of "spotted" had 20 spots, the 1/3 condition had 10 (mutatis mutandis for "dirty", which was generated by using clicks of the "spray can" function with grayish brown "paint").

- Finally, a 0/3 condition for the color quantity, color quality and minimum standard adjectives was simply an alien with zero degrees of the relevant property. But in the case of the relative adjectives "tall" and "wide", it doesn't make sense to refer

to an alien with zero degrees of height or width, so we created extremely short and extremely narrow aliens (both 1/2 the height or width of the 1/3 condition aliens) for the 0/3 relative adjective stimulus (see the first alien on the top left in Figure 8).

### 4.1.3 Experimental conditions

Experimental conditions were obtained by presenting two aliens, with different degrees of the relevant adjective. Hence, we refer to conditions with codes of the form "0/3 vs 3/3", here indicating that a 0/3 alien was presented with a 3/3 alien. We constructed the following experimental conditions for all adjectives:

- Three control conditions aimed to produce: clear cases of existence failure (0/3 vs 0/3), clear cases of uniqueness failure (3/3 vs 3/3), and clear cases of correct applications (0/3 vs 3/3).
- Three test conditions (0/3 vs 1/3, 1/3 vs 2/3, and 2/3 vs 3/3) aimed to evaluate under what conditions for each adjective participants would be willing or unwilling to accommodate existence and uniqueness presuppositions (see Table 3).

| **Control conditions** | | |
|---|---|---|
| 0/3 vs 0/3 | 3/3 vs 3/3 | 0/3 vs 3/3 |
| **NEITHER** | **BOTH** | **CORRECT** |
| *existence failure* | *uniqueness failure* | *clear correct answer* |

| **Target conditions** | | |
|---|---|---|
| 0/3 vs 1/3 | 1/3 vs 2/3 | 2/3 vs 3/3 |
| **CORRECT** | **CORRECT** | **CORRECT** |
| *relative + low threshold* | | |
| **NEITHER** | **CORRECT** | **CORRECT** |
| *relative + medium threshold* | | |
| **CORRECT** | **BOTH** | **BOTH** |
| *absolute + low threshold (and standard)* | | |
| **NEITHER** | **CORRECT** | **BOTH** |
| *absolute + medium threshold (and standard)* | | |
| **NEITHER** | **NEITHER** | **CORRECT** |
| *high threshold* | | |

Table 3: Predicted patterns of responses to the different conditions. The patterns are the same for all the control conditions, and different patterns in the target conditions identify different types of adjectives ("threshold" in this and subsequent tables picks out a *lower threshold* in the terminology of §1.2)

The total number of non-training items that participants responded to was 144: ((4 color quality target words + 4 color quantity target words + 2 relative target words +2 minimum standard target words) x 6 conditions = 72) x 2 (the alien stimuli were presented in switched positions (left vs. right) to control for order effects). All of the control and test items were presented in random order to all participants.

### 4.1.4 Response coding

Responses to the task were coded as follows:

| | |
|---|---|
| **CORRECT:** | Clicking on the alien with *more* of the relevant property |
| **INCORRECT:** | Clicking on the alien with *less* of the relevant property |
| **WHATEVER:** | Clicking on either of the aliens when they are identical |
| **NEITHER:** | Clicking on the "neither" button |
| **BOTH:** | Clicking on the "both" button |

## 4.2 Participants

We recruited 42 participants over Amazon Mechanical Turk for $0.80 each. One participant did not report English as their native language, and was excluded from our analyses. 17 participants were female, 22 male, and 2 other. Ages ranged from 24 to 66, and all participants correctly responded to a colorblindness test on the information form.

## 4.3 Results: Controls in the presupposition assessment task

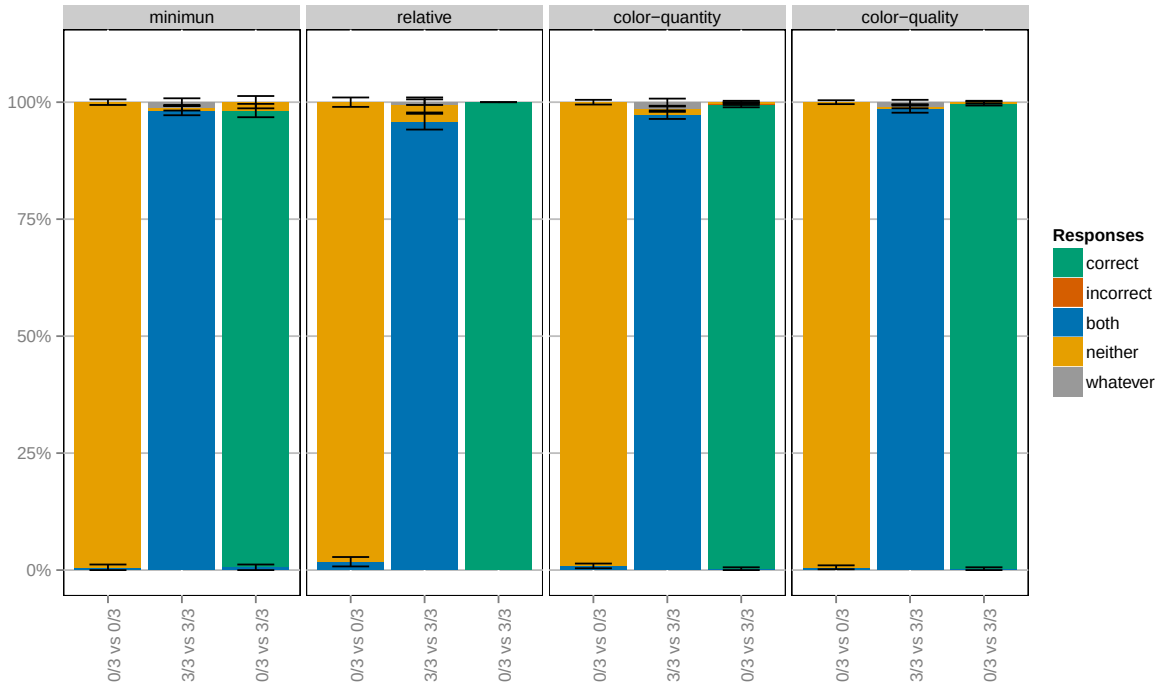



Figure 9: Mean percentage of each response types for the control conditions in the presupposition assessment task

Consider the three control conditions represented in Figure 9: The 0/3 vs 0/3 condition is a clear case of existence failure, the 3/3 vs 3/3 condition is a clear case of uniqueness failure, and in condition 0/3 vs 3/3 there is a clear correct response to the request. As is

evident from the stacked bar graph, participants are performing at or near ceiling with the control items (expected responses at least 95% of the time in each control condition for each type of adjective).

In the 0/3 vs 0/3 condition, participants almost universally responded with the response "neither are!", indicating existence failure. In the 3/3 vs 3/3 condition participants responded to the request to, e.g., *click on the red alien* when confronted with two equally, completely red aliens by responding with "both are!". And in the 0/3 vs 3/3 condition, where there is a clear correct response, subjects nearly universally responded with the "correct" response—that is, they picked the alien that had more of the relevant property. The expected responses hold for all adjective types.

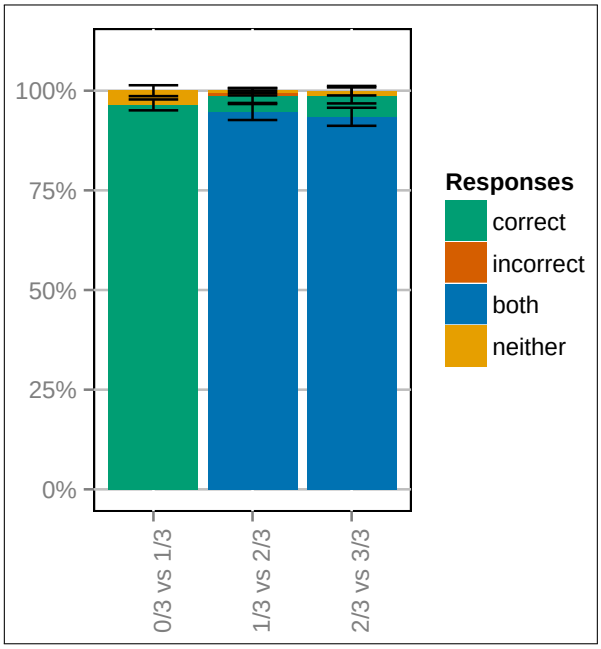
While responding correctly to these items is easy, the control condition results indicate that participants were paying attention and performing correctly throughout the experiment, because 72/144 of the experimental items that subjects responded to were controls, distributed randomly throughout the experiment.

### 4.4 Results: Minimum standard and relative adjectives

There are clear differences between responses to paradigmatically minimum standard and relative adjectives across all three test conditions (0/3 vs 1/3, 1/3 vs 2/3 and 2/3 vs 3/3). First, consider the chart in Figure 10, which represents responses to minimum standard adjectives across all three conditions. Responses display a distinctive pattern, which is what the standard theory predicts for minimum standard adjectives: subjects are choosing the alien with *more* of the relevant property only in the 0/3 vs 1/3 condition (M=96%), and then overwhelmingly rejecting the request to click on the alien with more of the relevant property in the 1/3 vs 2/3 and 2/3 vs 3/3 conditions (95% and 93%, respectively). The table in Figure 10 reveals that all participants choose that distinctive pattern (for each condition, we consider a preference for one of the four possible responses as unambiguous if a participant chose it at least 40% of the time in that condition).

Now consider the pattern of responses to relative adjectives represented in Figure 11. The first important result is that this pattern of responses is significantly different from the pattern of responses to minimum standard adjectives in all three conditions (e.g., whether we compare the amount of CORRECT responses or the amount of NEITHER responses (to apply a logit model), all $p$-values are below $.005$). That confirms the findings in Syrett et al. (2010). Focusing on the 2/3 vs 3/3 condition (on the far right of the bar chart in Figure 11), subjects responded to relative adjectives overwhelmingly (M=99.4%) by picking the alien with more of the relevant property (more height, more width). In contrast, the overwhelming mean response to minimum standard adjectives in this condition was to refrain from picking the alien with more of the relevant property and respond with "both are!" (a refusal to accommodate the uniqueness presupposition of the definite description) on average 94.5% of the time.

Responses to relative adjectives in the 0/3 vs 1/3 and 1/3 vs 2/3 conditions are puzzling at first glance, because there is some amount of "neither are!" response in both conditions (M=52% and 35%, respectively), which should only characterize *maximum standard* adjec-




| 0/3 vs 1/3 | 1/3 vs 2/3 | 2/3 vs 3/3 | Count |
|---|---|---|---|
| **CORRECT** | **BOTH** | **BOTH** | 42 |

*absolute + low threshold (and standard)*

Figure 10: Responses for minimum standard adjectives, (a) in the population, (b) counts of patterns of responses for individuals (where individuals are classified as having given a particular response when they give it at least 40% of the time), with an interpretation for the given pattern of responses, when possible.

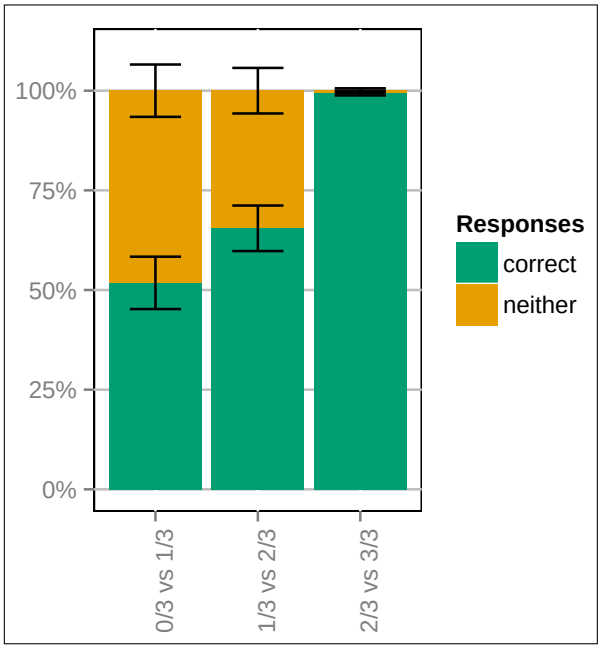
tives on the standard view. Syrett et al. (2010, p. 11) predict that participants will always be able to accommodate the uniqueness and existence presuppositions of definite descriptions when combined with relative adjectives (as in "Please click on the tall alien"):

> Because relative GAs [gradable/scalar adjectives] such as 'big' and 'long' depend on the context for the standard of comparison, participants should posit a new standard of comparison each time a new pair is introduced in order to ensure that the adjective is true of just one object (i.e. the bigger or longer one). Thus, participants should always be able to accommodate the presuppositions of the definite description and accept the request as felicitous.

But our results indicate that a significant number of subjects don't accommodate with relative adjectives that way.

What is going on with relative adjectives? While the observed pattern of responses conflicts with the predictions of the "off the shelf" picture, it is compatible with the picture of hybrid standards described in §1.2, above. According to the alternative picture, an object has to meet or exceed the lower threshold before some participants are willing to accommodate the existence presupposition of the definite description.[18] The pattern of responses to

[18] Syrett et al. (2010) and Kennedy (2007) discuss what they call a "threshold effect" that might initially seem like a plausible candidate to explain the rejection of the existence presupposition. The "threshold




| 0/3 vs 1/3 | 1/3 vs 2/3 | 2/3 vs 3/3 | count |
|---|---|---|---|
| **CORRECT** | **CORRECT** | **CORRECT** | 29 |
| *relative + low threshold* | | | |
| **NEITHER** | **CORRECT** | **CORRECT** | 7 |
| *relative + medium threshold* | | | |
| **NEITHER** | **NEITHER** | **CORRECT** | 6 |
| *high threshold* | | | |

Figure 11: Responses for relative adjectives, see Figure 10 for details.

relative adjectives that we observe is evidence not only of the existence of such a threshold, but that there is some interpersonal variation in where on the scale that threshold is located.

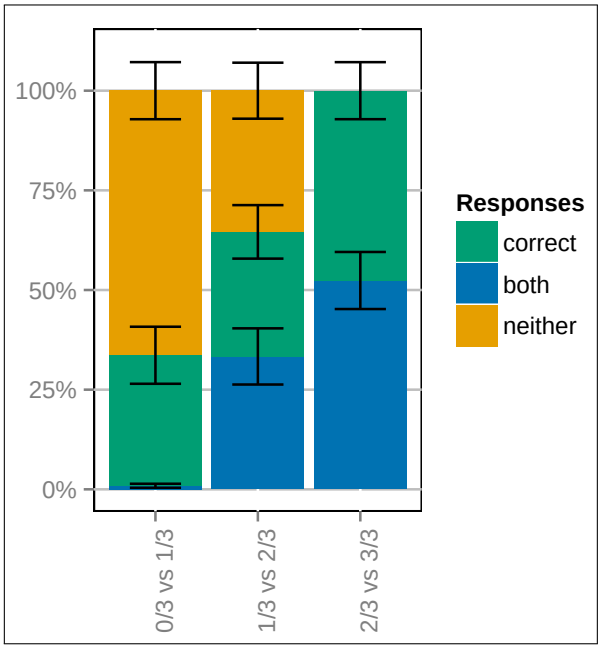
### 4.5 Results: Color Quantity

We now turn to consider how participants respond to the quantitative reading of color adjectives, represented in Figure 12.

First of all, the pattern of responses to the quantitative reading of color adjectives is significantly different than either the pattern observed for minimum standard or relative adjectives.[19] What's going on? Do different participants respond to color quantity adjec-

effect" shows up as the unwillingness of speakers to apply a relative adjective to an object that has a small, but noticeablly greater degree of the relevant property. So, for example, if I ask you to "Click on the tall alien" when one alien is only slightly taller than the other, you would refuse (according to the account in Syrett et al. 2010 and Kennedy 2007). The threshold effect is due, according to Syrett et al., to the underlying vagueness of relative adjectives, and an unwillingness to distinguish objects that are "very similar to each other relative to the scalar property that the adjective encodes" (see also Kennedy 2007, pp. 18–19). (This is the unwillingness to make crisp distinctions that drives the sorites paradox.) If this "threshold effect" due to vagueness explains the failures to accommodate in the 0/3 vs 1/3 and 1/3 vs 2/3 conditions, then it should generate similar failures to accommodate in the 2/3 vs 3/3 condition, since the heights of the stimuli vary regularly across conditions. (In fact, the 0/3 condition is 1/2 of the height of the 1/3 condition. So there's an even greater difference in height in the 0/3 vs 1/3 condition than there is in the 1/3 vs 2/3 and 2/3 vs 3/3 conditions. That should make it *easier* to accommodate in the 0/3 vs 1/3 condition.) But participants don't fail to accommodate in the 2/3 vs 3/3 condition, so the Syrett et al. and Kenney "threshold" effect can't be the explanation for the failures to accommodate in the 0/3 vs 1/3 and 1/3 vs 2/3 conditions.

[19]We can see for instance that, according to our usual logit models, the proportion of NEITHER responses is higher for color than for relative adjectives both in 0/3 vs 1/3 and in 1/3 vs 2/3 conditions ($p < .001$) and




| 0/3 vs 1/3 | 1/3 vs 2/3 | 2/3 vs 3/3 | count |
|---|---|---|---|
| **CORRECT** | **BOTH** | **BOTH** | 13 |
| *absolute + low threshold (and standard)* | | | |
| **NEITHER** | **NEITHER** | **CORRECT** | 10 |
| *high threshold* | | | |
| **NEITHER** | **CORRECT** | **BOTH** | 9 |
| *absolute + medium threshold (and standard)* | | | |
| **NEITHER** | **CORRECT** | **CORRECT** | 3 |
| *relative + medium threshold* | | | |
| **BOTH** | **BOTH** | **BOTH** | 1 |
| **BOTH** | **CORRECT** | **CORRECT** | 1 |
| **BOTH** | ? | **BOTH** | 1 |
| **CORRECT** | CORRECT | **BOTH** | 1 |
| **NEITHER** | **BOTH** | **BOTH** | 1 |
| **NEITHER** | **BOTH** | **CORRECT** | 1 |
| **NEITHER** | **NEITHER** | **BOTH** | 1 |

Figure 12: Responses for the quantitative readings of color adjectives. The question mark indicates a failure to choose a response consistently (more than 40% of the time in the relevant condition)

tives as if they were minimum standard and others respond to them as if they were relative? Are participants responding inconsistently? Do color quantity adjectives break the standard mold for classifying scalar adjectives?

By looking at individual responses in the table of Figure 12, we get a more fine-grained picture of how the quantitative reading of color adjectives relates to relative and minimum standard adjectives. Responses to the quantitative reading of color adjectives fall mainly into three patterns: either an absolute + low threshold and standard (**CORRECT-BOTH-BOTH**) pattern, an absolute + high threshold and standard pattern (**NEITHER-NEITHER-CORRECT**), and an absolute + medium threshold and standard (**NEITHER–CORRECT–BOTH**) pattern. The absolute + medium threshold and standard pattern conforms with the account of the quantitative reading of color adjectives given in McNally (2011), in which for something to count as having a certain color, that color has to "predominate", but once the color predominates, the color adjective behaves like an absolute adjective. These three dominant patterns are followed by a motley tail of responses that don't clearly align with any standard.[20]

that the proportion of BOTH responses is higher for color than for minimum standard adjectives in the 1/3 vs 2/3 condition ($p < .001$), and there are even more such BOTH responses in the 2/3 vs 3/3 condition for color adjectives (but not for minimum standard).

[20]To assess which set of patterns are significantly populated by participants beyond chance, one can run successive $\chi^2$-tests: with the whole set of (observed) patterns first, and then dropping the next most extreme remaining pattern, one after the other. Supposing that the range of possible patterns is the one we ended up observing (which is a conservative hypothesis because there were much more possible patterns, which means

This variation indicates that while both of the existing hypotheses concerning the meaning of the quantitative reading of color adjectives (Clapp's minimum standard hypothesis and McNally's absolute + medium standard hypothesis) are present in some subjects' responses, neither of those hypotheses fully captures the variety of how subjects respond to the quantitative reading of color adjectives.

### 4.6 Results: Color quality

Responses to the qualitative reading of color adjectives are represented in Figure 13. While responses to the qualitative reading of color adjectives clearly differ from responses to both minimum standard and relative adjectives, discerning a clear pattern within responses to the qualitative reading is more difficult.[21]

Looking at the counts of participants responding with different patterns, there is a majority absolute + low threshold and standard response pattern (**CORRECT-BOTH-BOTH**), followed by no clear pattern of responses.[22] The "?" response, when it appears throughout the table in Figure 13, indicates that the relevant participants did not choose any particular response more than 40% of the time: more than a third of the participants (16 out of 42) were affected.

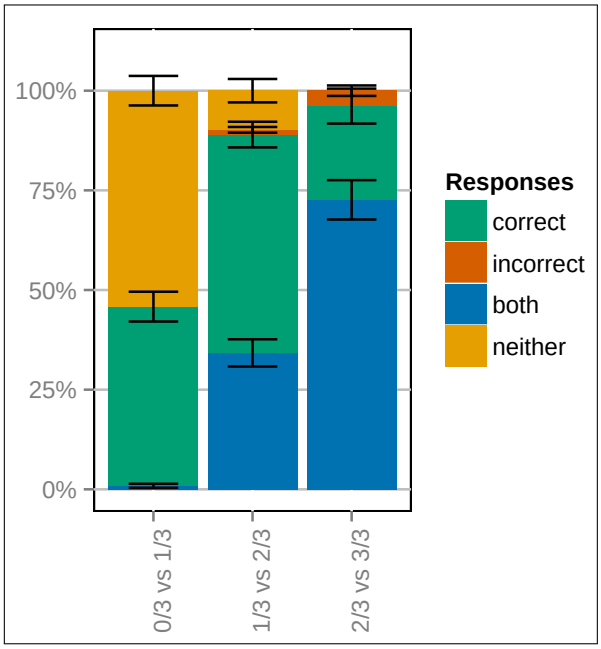
that we expect lower extremes), these tests tell us at each stage if the maximal extreme value that remains in the set does contribute a significant divergence from chance. The results are as in the table below, showing that the first three extreme patterns, with 13, 10 and 9 participants respectively, contain more participants than expected by chance. The next pattern, with 3 participants, does not deviate from chance. For this to reach significance, one would need to assume that there are 34 unobserved possible patterns (or more), and even then we would only reach the .05 significance threshold, which is not sufficient if we take into account the need for correction for multiple comparisons.

| distribution | $\chi^2$ | $p$ |
|---|---|---|
| 13,10,9,3,1,1,1,1,1,1,1 | 54 | $5.10^{-8}$ |
| 10,9,3,1,1,1,1,1,1,1 | 39 | $1.10^{-5}$ |
| 9,3,1,1,1,1,1,1,1 | 27 | $7.10^{-4}$ |
| 3,1,1,1,1,1,1,1 | 2.8 | .90 |
| 3,1,1,1,1,1,1,1 adding 34 empty cells (with 0s) | 41 | .048 |

[21]For instance, color adjectives generate more 'neither' responses than minimum standard adjectives in the 0/3 vs 1/3 condition ($p < 1.10^{-14}$) and more 'both' responses than relative adjectives in the 1/3 vs 2/3 condition ($p < 1.10^{-14}$).

[22]As argued in footnote 20, the following tests show that only the first pattern (with 22 participants) is unambiguously endorsed by more participants than what is expected by chance. It is also worth noting that the second pattern is not really unambiguously given that it is made of participants for which no clear response choice emerged in the 1/3 vs 2/3 condition.

| distribution | $\chi^2$ | $p$ |
|---|---|---|
| 22,6,3,2,2,2,1,1,1,1,1 | 101 | $2.10^{-16}$ |
| 6,3,2,2,2,1,1,1,1,1 | 11 | .28 |
| 6,3,2,2,2,1,1,1,1,1 adding 6 empty cells (with 0s) | 15 | .013 |




| 0/3 vs 1/3 | 1/3 vs 2/3 | 2/3 vs 3/3 | count |
|---|---|---|---|
| **CORRECT** | **BOTH** | **BOTH** | 22 |
| *absolute + low threshold (and standard)* | | | |
| **CORRECT** | ? | **BOTH** | 6 |
| **CORRECT** | ? | ? | 3 |
| **CORRECT** | **BOTH** | ? | 2 |
| **CORRECT** | **CORRECT** | **BOTH** | 2 |
| ? | **BOTH** | **BOTH** | 2 |
| **CORRECT** | **CORRECT** | **CORRECT** | 1 |
| ? | ? | ? | 1 |
| **NEITHER** | **CORRECT** | **BOTH** | 1 |
| **NEITHER** | **CORRECT** | ? | 1 |
| **NEITHER** | ? | **BOTH** | 1 |

Figure 13: Responses for the qualitative readings of color adjectives, see Figure 10 for details. Question marks indicate a failure to choose a response consistently (more than 40% of the time in the relevant condition)

### 4.7 Discussion

In the application of the presupposition assessment task to both readings of color adjectives, we have replicated the sharp difference Syrett et al. (2010) found when the test is applied to what are standardly regarded as minimum standard absolute and relative adjectives. But we also found evidence of lower thresholds for relative adjectives that objects need to meet or exceed before some subjects are willing to accommodate the existence presuppositions of definite descriptions, contrary to the predictions in Syrett et al. (2010).

We found evidence of three distinct ways of understanding the standards associated with the quantitative reading of color adjectives: absolute + low threshold and standard, high threshold (which might also be evidence of a maximum standard absolute interpretation), and absolute + medium threshold and standard. Existing hypotheses about the meaning of the quantitative reading of color adjectives do not predict this variety.

The qualitative reading of color adjectives, on the other hand, did not produce any clear pattern of responses. This could be due either to non-semantic or semantic factors. The non-semantic factors would include subjective variation in how participants perceive color, or the conditions in which the experiment was conducted (we can't control features of the screen or lighting conditions in which online participants perform the task). Alternatively, the variation we observed could be due to variation in how participants interpret a multidimensional property like color quality, or in terms of variation in where participants locate thresholds on the scale. Or some combination of all of these factors.

Future experiments could distinguish between these possibilities. For example, if par-

ticipants responded similarly to other types of multidimensional adjectives ("sick", "healthy", "beautiful", "ugly") that aren't susceptible to the same kinds of perceptual and environmental variation, that would be evidence that the variation in the qualitative reading of color adjectives isn't just due to perceptual or environmental factors.[23]

### 4.7.1 Possible concerns

First, one might wonder whether we chose the wrong examples of the 2/3 vs 3/3 condition for qualitative reading of color adjectives. After all, these involve potentially idiosyncratic subjective judgments of what counts as the "best" example of the relevant color. But it turns out that this potential idiosyncracy does not matter, because (a) we could potentially count either a **CORRECT** or an **INCORRECT** response as indicative of a relative type of response, and (b) there there were very few **INCORRECT** responses anyway (see Figure 13), indicating that almost all participants agreed with our qualitative orderings.

Second, a defender of the traditional view of standards might object that what we take to be evidence of the existence of thresholds could be explained instead by the fact that participants might not just be comparing the two aliens on the screen, but also the aliens they have recently seen in the experiment. As participants see more examples of aliens, including examples that are taller than the ones on the screen they are currently viewing, they will be reluctant to pick one of the two non-maximally tall aliens on the screen in response to the request to "Click on the tall alien". That would then explain the tendency of some participants to respond to the request with "Neither is!" when neither alien is maximally tall, without the need to introduce the idea of thresholds.[24]

One way to test this possibility is to see if this response (the "Neither is!" response to relative adjectives in the 0/3 vs 1/3 and 1/3 vs 2/3 conditions) becomes more frequent the further into the experiment participants get. If it does, that would be evidence in support of the idea that participants weren't just comparing the aliens on the screen in front of them, but had other aliens that they had seen during the experiment in mind.

On one hand, there is some indication of such an effect: the latter in the experiment, the more "neither" responses appear (2/3 vs 3/3: $p = .049$, but 1/3 vs 2/3: $p = .20$). On the other hand, there is a similar effect as the experiment goes on for, e.g., "both" responses for minimum standard adjectives in the 2/3 vs 3/3 condition ($p < .001$). That suggests that participants are simply performing the task differently by the end of the experiment, which makes the order effects on "neither" responses difficult to interpret.

### 4.7.2 Relation to the first experiment

What do the results of the entailment experiment tell us about the existence of hybrid standards? If the standard is located anywhere other than the minimal degree on the scale, then the entailments that characterize minimum standard adjectives no longer hold. That

[23] See Sassoon (2013) and McNally and Stojanovic (2014) for discussion of multidimensional adjectives.

[24] James Hampton and Robert van Rooij made this suggestion in discussion.

includes McNally's intermediate absolute standard for the quantitative reading of color adjectives and the "high threshold" standard, both of which we found some evidence of in the presupposition accommodation experiment.

If, e.g., some people interpret the quantitative reading of color adjectives as having a standard around the midpoint of the scale, then they should not be willing to infer "X is red" from "X is redder than Y". Given that, we might expect to see different results on the inference tests than we in fact found. Namely, responses to color adjectives should differ from responses to paradigm minimum standard adjectives like "spotted". But we didn't observe such a difference. Why not?

One possibility is that participants are suffering from an understandable failure of imagination when they engage in the inference tests. In order to detect that, e.g., "X is redder than Y" does *not* entail "X is red" (if the standard is somewhere around the midpoint of the scale), participants would need to imagine two things with, e.g., small amounts of red on them. That failure to imagine some relevant possibilities would make color terms look like they have standards at scale minima when in fact they don't. A future experiment could evaluate this possibility, by looking at the results of the inference tests after participants are primed with examples of objects that have some degree of redness, but a degree far below the midpoint.

## 5 Conclusions and further research

One major advantage of looking at context sensitivity through the lens of scalar adjectives is that scalar adjectives have been closely studied by linguists, and that distinctions between types and degrees of context sensitivity applying to adjectives are fine-grained. Debates about the philosophical significance of context sensitivity can thus be anchored to a substantial foundation of linguistic data and theory.

Furthermore, the advantage of investigating the nature of standards for different types of adjectives using a formal experimental approach is that it reveals that various existing accounts of the standards appropriate for color adjectives are all only partially correct. It turns out that the quantitative reading involves interpersonal variation about where the standard is located: some participants treat the quantitative reading as minimum standard-like (in alignment with Clapp's prediction), some treat it as having a very high threshold (or possibly as maximum standard absolute), and other participants treat it as somewhere in between (in accordance with McNally's prediction). The qualitative reading, on the other hand, displays no clear pattern of responses beyond a majority minimum standard response. The explanation for the scattered responses isn't yet clear, but there are two plausible possibilities to investigate in further research: that the variation is due to non-semantic factors (perception, for example), or to semantic factors (multidimensionality).

Most interestingly, our version of the presupposition accommodation task uncovered evidence that lends support to the hybrid picture of the nature of standards outlined in §1.2. We found evidence of a lower threshold in the fact that some participants respond to paradigmatically relative adjectives like "tall" by refusing to click on the *taller* of the two objects when they fall below a certain threshold of height. More sensitive experimental

studies could probe for evidence of the upper threshold in paradigmatically relative adjectives as well, by giving participants examples of objects that both clearly have a relevant property (two extremely tall sequoias, for example), but that clearly differ in height, and asking participants to click on "the tall one". And a finer-grained experimental approach could look for evidence of the strong view about hybrid standards (that all adjectives are hybrid to some degree) by looking at paradigmatically minimum standard and maximum standard adjectives to see whether they behave like relative adjectives when applied to objects that have very low or very high degrees on the relevant scale. That is work for a future study.

The significance of these experimental findings for the debate about context sensitivity, which prompted this investigation, cuts in different directions. On one hand, neither the quantitative reading nor the qualitative reading behaves like paradigmatically relative adjectives, so the context sensitivity they display is not obviously due to their standards shifting in different contexts in the way that the standard associated with, e.g., "tall" does. On the other hand, the quantitative reading displays significant interpersonal variation in where the standard is located on the scale, and the qualitative reading displays a great deal of variation, both between and within subjects. The existence of those forms of variability in how we understand color adjectives needs to be part of the debate about how context and meaning interact.

## References


Bartsch, R. and Vennemann, T. (1972). The grammar of relative adjectives and comparison. *Linguistische Berichte*, 20:19–32.

Burnett, H. (2012). The puzzle(s) of absolute adjectives. *UCLA Working Papers in Linguistics, Papers in Semantics*, 16:1–50.

Cappelen, H. (2012). *Philosophy without Intuitions*. Oxford University Press, Oxford.

Clapp, L. (2012). Indexical color-predicates: Truth-conditional semantics vs. truth-conditional pragmatics. *Canadian Journal of Philosophy*, 42(2):71–100.

Davies, A. (2015). Off-target responses to occasion-sensitivity. *dialectica*.

DeRose, K. (2008). Gradable adjectives: A defense of pluralism. *Australasian Journal of Philosophy*, 86(1):141–160.

Glanzberg, M. (2007). Context, content, and relativism. *Philosophical Studies*, 136(1):1–29.

Hansen, N. (2011). Color adjectives and radical contextualism. *Linguistics and Philosophy*, 34(3):201–221.

Hansen, N. and Chemla, E. (2013). Experimenting on contextualism. *Mind & Language*, 28(3):286–321.

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

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

Kennedy, C. and McNally, L. (2010). Color, context, and compositionality. *Synthese*, 174(1):79–98.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. *Proceedings of SALT*, 20:197–215.

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In Nouwen, R., van Rooij, R., Sauerland, U., and Schmitz, H.-C., editors, *ViC 2009: Papers from the ESSLLI 2009 Workshop on Vagueness in Communication*, volume 6517 of *Lecture Notes in Computer Science*, pages 151–168, Heidelberg. Springer-Verlag.

McNally, L. and Stojanovic, I. (2014). Aesthetic adjectives. In Young, J., editor, *Semantics of Aesthetic Judgement*, Oxford. Oxford University Press.

Rothschild, D. and Segal, G. (2009). Indexical predicates. *Mind & Language*, 24(4):467–493.

Rotstein, C. and Winter, Y. (2004). Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics*, 12(3):259–288.

Sassoon, G. W. (2011). A *Slightly* modified economy principle: Stable properties have non-stable standards. Talk at Workshop on Degree Semantics and its Interfaces, Utrecht.

Sassoon, G. W. (2013). A typology of multidimensional adjectives. *Journal of Semantics*, 30:335–380.

Solt, S. (2011). Comparison to arbitrary standards. *Sinn und Bedeutung*, 16:1–14.

Stanley, J. (2004). On the linguistic basis of contextualism. *Philosophical Studies*, 119(1–2):119–146.

Syrett, K., Kennedy, C., and Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, 27(1):1–35.

Syrett, K. L. (2007). *Learning about the Structure of Scales: Adverbial Modification and the Acquisition of the Semantics of Gradable Adjectives*. PhD thesis, Northwestern University, Evanston, Illinois.

Szabó, Z. G. (2001). Adjectives in context. In Kenesei, I. and Harnish, R. M., editors, *Perspectives on Semantics, Pragmatics, and Discourse: A Festschrift for Ferenc Kiefer*, pages 119–146. John Benjamins Publishing Company, Amsterdam.

Toledo, A. and Sassoon, G. W. (2011). Absolute vs. relative adjectives – varience within vs. between individuals. In *Proceedings of SALT 21*, pages 135–154, Rutgers University. MIT Working Papers in Linguistics.

Travis, C. (2008a). *Occasion-Sensitivity: Selected Essays*. Oxford University Press, Oxford.

Travis, C. (2008b). Pragmatics. In *Occasion-Sensitivity: Selected Essays*, pages 109–129. Oxford University Press, Oxford.

Unger, P. (1975). *Ignorance: A Case for Skepticism*. Oxford University Press, Oxford.

Vicente, A. (2015). The green leaves and the expert: Polysemy and truth-conditional variability. *Lingua*.

Yoon, Y. (1996). Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*, 4:217–236.