Color adjectives, standards, and thresholds: an experimental investigation

Linguist and Philos
DOI 10.1007/s10988-016-9202-7

Color adjectives, standards, and thresholds: an experimental investigation

Nat Hansen[1,2] · **Emmanuel Chemla**[3]

Abstract Are color adjectives ("red", "green", etc.) relative adjectives or absolute adjectives? Existing theories of the meaning of color adjectives attempt to answer that question using informal ("armchair") judgments. The informal judgments of theorists conflict: it has been proposed that color adjectives are absolute with standards anchored at the minimum degree on the scale, that they are absolute but have near-midpoint standards, and that they are relative. In this paper we report two experiments, one based on entailment patterns and one based on presupposition accommodation, that investigate the meaning of scalar adjectives. We find evidence confirming the existence of subgroups of the population who operate with different standards for color adjectives. The evidence of interpersonal variation in where standards are located on the relevant scale and how those standards can be adjusted indicates that the existing theories of the meaning of color adjectives are at best only partially correct. We also find evidence that paradigmatic relative adjectives ("tall", "wide") behave in ways that are not predicted by the standard theory of scalar adjectives. We discuss several different possible explanations for this unexpected behavior. We conclude by discussing the relevance of our findings for philosophical debates about the nature and extent of semantically encoded context sensitivity in which color adjectives have played a key role.

✉ Nat Hansen
n.d.hansen@reading.ac.uk

Emmanuel Chemla
chemla@ens.fr

[1] Department of Philosophy, University of Reading, Reading RG6 6AA, UK

[2] Humanities Center, Stanford University, Stanford, CA, USA

[3] Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, École Normale Supérieure, PSL Research University, Paris, France

Published online: 10 March 2017

Keywords Scalar adjectives · Color terms · Context · Contextualism · Standards · Experimental semantics

1 Introduction

Scalar (or gradable) adjectives denote properties that can apply to objects to varying degrees: one lunch can be *longer* than another; one book can be *more expensive* than another. A standard semantics distinguishes 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. duration or price). In contrast, *absolute* adjectives like “full” or “spotted” have conventionally fixed reference points, and therefore display less semantic context sensitivity than relative adjectives.

Given that color adjectives are a type of scalar adjective (an apple can be “redder than” another, “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.

There are several competing answers to that question that have been offered by theorists of color adjectives. Hansen (2011) focuses on the extensive contextual variability that color adjectives display, and assumes that color adjectives have relative standards. Clapp (2012) argues 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. And McNally (2011) proposes that color adjectives are absolute, but that they have a standard that lies somewhere near the midpoint of the relevant scale. All of these existing proposals rely exclusively on informal judgments as evidence.

In this paper, we put the debate about the meaning of color adjectives on a firmer empirical foundation by using two experimental methods of evaluating the semantic properties of scalar adjectives:

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

We find that when color adjectives are subjected to these tests, they display surprising patterns that diverge from existing armchair judgments: once we disambiguate two possible (*quantitative* and *qualitative*) readings of color adjectives (Kennedy and McNally 2010), we find that responses 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 an absolute standard were located somewhere in the middle of the scale, and a third that responds as if an absolute standard were located at the upper end of the scale. In contrast to 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 also found a surprising pattern of responses to paradigmatic relative adjectives: certain participants are reluctant, when evaluating objects that fall on the low end of the relevant scale (of tallness, for example), to accommodate the existence presupposition associated with the definite description in requests like "Please click on the tall alien". 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.

In Sect. 2, we outline the degree-based semantics for scalar adjectives that is the background for our experiments. Section 3 discusses an experiment that evaluates color adjectives alongside paradigm relative and minimum standard absolute adjectives using *entailment tests* that are diagnostic of different types of standards. Section 4 discusses an experiment that uses patterns of *presupposition accommodation* to distinguish different types of adjectives. Section 5 relates the results of our two experiments to philosophical debates about the significance of pervasive context sensitivity. Finally, Sect. 6 considers directions for further research.

2 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 the standard "off the shelf" semantics for scalar adjectives, they are associated with functions from individuals to degrees on a scale (Bartsch and Vennemann 1972; Kennedy 2007; Syrett et al. 2010):

(5) $[\![\textbf{prime}_{\langle e,t\rangle}]\!] = \lambda\text{x} \,.\, \text{prime(x)}$
(6) $[\![\textbf{expensive}_{\langle e,d\rangle}]\!] = \lambda\text{x} \,.\, \text{expensive(x)}$

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 if and only if the hardcover is mapped to a greater degree on the scale of cost than the paperback:

(7) $[\![\textbf{more G than}_{\langle\langle e,d\rangle,\langle e,\langle e,t\rangle\rangle\rangle}]\!] = \lambda\text{G}\lambda\text{y}\lambda\text{x} \,.\, \text{G(x)}\succ \text{G(y)}$
(8) $[\![\textbf{more expensive than}_{\langle\langle e,\langle e,t\rangle\rangle\rangle}]\!] = \lambda\text{y}\lambda\text{x} \,.\, \text{expensive(x)} \succ \text{expensive(y)}$
(9) $[\![\textbf{The hardcover is more expensive than the paperback}_{t}]\!] = \text{expensive(the hardcover)} \succ \text{expensive(the paperback)}$

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 provides the relevant comparison.[1] 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 a context-sensitive function *standard* to the scalar adjective it combines with. The standard supplied by *pos* "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) $[\![\textbf{pos}_{\langle\langle e,d\rangle,\langle e,t\rangle\rangle}]\!] = \lambda G \lambda x \,.\, G(x) \succeq \text{standard}(G)$
(13) $[\![\textbf{pos expensive}_{\langle e,t\rangle}]\!] = \lambda x \,.\, \text{expensive}(x) \succeq \text{standard}(\text{expensive})$
(14) $[\![\textbf{The hardcover is pos expensive}_{t}]\!] = \text{expensive}(\text{The hardcover}) \succeq \text{standard}(\text{expensive})$

So (14) is true if and only if 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 bindings 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 nice 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. Galileo, in his *Dialogue Concerning the Two Chief World Systems*, writes:

> [T]hese 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'.[2]

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 been divided into two further 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

[1] For an extended critical discussion of the role of the null degree morpheme *pos*, see Rett (2015).

[2] http://goo.gl/NSF6CD.

associated with the adjective. 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.[3]

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) “x is F_{min}” is true if and only if x has more than a zero degree of F-ness
(16) “x is F_{max}” is true if and only if x has the maximum degree on the scale of F-ness
(17) “x is F_{rel}” is true if and only if x has a degree on the scale of F-ness that is equal to or greater than the contextually determined standard

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.

So how do color adjectives fit into the standard typology of scalar adjectives? In order to answer that question, we conducted two experiments, one based on the distinctive entailment patterns in which different types of scalar adjectives appear, and one based on Syrett et al.’s (2010) presupposition assessment task, which reveals differences in speakers’ willingness to accommodate the presuppositions of definite descriptions when combined with different types of adjectives.

3 Experiment #1: Evidence from entailment patterns

One standard way of distinguishing relative from absolute adjectives is in terms of the entailments they license. 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” if 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 typically have more than a minimal degree of the relevant property to count as satisfying the adjective. Being tall, for example, requires more than just having a minimal degree of height.[4]

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; van Rooij 2009):

[3] For criticism of the standard way of drawing the relative-absolute distinction and alternative theories, see Burnett (2014), Klein (1980) and Toledo and Sassoon (2011).

[4] Panzeri and Foppolo (2012), however, report an experiment that seems to indicate that children (3- and 5-year olds) and adults who have been instructed to be extremely conversationally tolerant tend to find applications of relative scalar adjectives to an object acceptable if the object has *some degree* of the relevant property. Thanks to an anonymous referee for the reference.

(18) **Entailment pattern #1**
X is more Adj than y ⇒ X is Adj.

That is, if x has more of the property denoted by "Adj" than y does, then x has some degree of the property denoted by "Adj". On the standard picture of minimum standard absolute adjectives, x is only required to have a minimal degree of Adj-ness to satisfy the adjective, so x is Adj.[5]

(19) **Entailment pattern #2**
X is not Adj ⇒ x has a zero degree of Adj-ness.

That is, if all it takes for x to have the property denoted by "Adj" is to have a minimal degree of Adj-ness, then not having that property means x has zero 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!

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) test existing informal judgments about the different entailment patterns that relative and minimum standard absolute adjectives are supposed to license, and (2) determine whether or not color adjectives display similar patterns. Note that both informal 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 interfering.

3.1 Participants

41 participants were recruited online for experiment 1 using Amazon Mechanical Turk, and were paid $0.80 each for participation in the experiment.[6] Participants were required to have U.S. IP addresses, to have at least 95% approval rate for previous HITs, were recruited as a single group, assigned either to experiment 1 or experiment 2 (discussed below), and there was no overlap in participants between the two experiments. One participant did not report being a native English speaker and was therefore excluded from the analyses. Ages ranged from 21 to 66. 19 participants were female, and 22 male.

[5] 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.

[6] See Sprouse (2011) for a discussion of the use of Mechanical Turk for running linguistic experiments.

Table 1 Target adjectives in the entailment experiment

Minimum standard	Bumpy	Dirty	Sick	Spotted	Visible	Wet
Relative	Big	Heavy	Long	Old	Tall	Wide
Color	Blue	Brown	Green	Pink	Red	Yellow

3.2 Materials, design and task

3.2.1 Adjectives

We tested six adjectives of each of three types: minimum standard absolute, relative, and color (see Table 1). The examples of the relative and minimum standard absolute adjectives were chosen because they commonly appeared in lists illustrating each type of adjective in the literature, with a preference for simple, single-dimensional adjectives ("long", e.g.) over more complex multi-dimensional adjectives ("intelligent", e.g.). The color terms are a mix of basic color terms and two non-basic terms ("pink", "brown").[7]

3.2.2 Three inferential tasks

To test the entailment patterns that different types of adjectives license, we used three different inferential tasks for each of the two entailment patterns discussed above in (18) and (19).

- The downward arrow task ("↓") is intended to elicit more or less direct judgments about entailment. After a brief introduction to entailment (see Appendix A for details), 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 (20a).
- 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* being the appropriate premise and conclusion as in (20b).
- 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, as in (20c); negative responses here indicate entailment from *p* to *q*.

(20) The following illustrates the three inferential tasks, featuring the adjective 'tall' and the first entailment pattern [see (18)]:

a. Downward arrow task "↓":

"X1 is taller than Y2."
↓
"X1 is tall."

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

[7] For discussions of single- versus multi-dimensional adjectives, see Sassoon (2012) and Solt (2016).

Table 2 Predictions for the inferential tasks

	Relative	Minimum	Color
↓	No	Yes	?
THEREFORE	No	Yes	?
BUT	Yes	No	?

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

3.2.3 Instructions and order of presentation

Each participant was presented with 108 test items, which were the result of combining 3 adjective types × 6 adjectives × 2 entailment patterns × 3 inferential tasks. 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 instructions and practice items (which were included to let participants get used to the display, but which weren't testing our target expressions), and participants were randomly assigned to either a "↓"-first or "↓"-second ordering of the blocks.[8] We observed no relevant order effects of blocks.[9]

3.2.4 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.

3.3 Results of the entailment pattern experiment

Figure 1 reports mean proportion of "YES" responses to the test items. The first result which emerges is that, as expected, responses to the "but" inference task are the mirror image of the responses in the "therefore"/"↓" tasks. This indicates that participants were tracking the requirements of the different inference tasks. Formally, if we consider the mean responses in each condition, the pairwise Pearson correlation cœfficients between the tasks are 98% (↓ vs. "therefore"), −97% (↓ vs. "but") and

[8] 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.

[9] This is based on visual inspection. More formally, we can assess this question using binomial mixed model comparisons as below (here solely with intercept random effects for participants and items, since models with a full random structure did not converge). The main model would have adjective type, inference type and inferential task as fixed factors, together with the interactions of each of these with the order factor. This full model is not significantly superior to one in which the interaction between order and adjective type is dropped ($\chi^2(2) = 2.78$, $p = .25$), which suggests that the effect of adjective type, which is our main interest, is not affected by order.

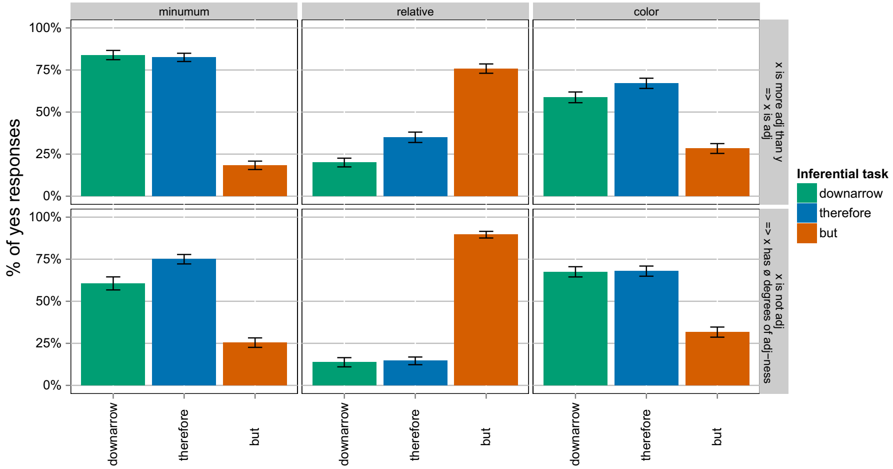


Fig. 1 Percentages of "Yes" responses to the entailment pattern experiment

−99% ("therefore" vs. "but"). Similarly, the two inference patterns provide the same information (correlation coefficient is 96%).

In a second, more interesting, analysis we compared adjective types two by two. To do so, we first reversed the value of the responses to the "but" task, so that all tests would be aligned (a high score on any of the tests now means that the inference goes through, for all three tests). We then fit the data (restricted to the adjective types of interest) with a generalized linear binomial mixed model, with adjective type as a fixed effect, as well as inference type and inferential task as fixed effects to control for their potential remaining effects. When possible, we used a maximal random structure for participants and item (i.e. actual adjective) (see Barr et al. 2013). We used the lme4 package from *R* to do so (Team 2015; Bates et al. 2015). To evaluate the role of the adjective type, we compared such a model with an identical one, which gives no role to adjective type.

Such model comparison analyses revealed a significant difference between minimum and relative adjectives: ($\chi^2(1) = 34$, $p < 5 \times 10^{-9}$)[10] a significant difference between color and relative adjectives ($\chi^2(1) = 30$, $p < 5 \times 10^{-8}$) and no significant difference between color and minimum standard absolute adjectives ($\chi^2(1) = 1.7$, $p = .20$).[11] In summary, the inferences we tested go through with minimal adjectives but not with relative adjectives, as predicted in the literature, and they also go

[10] The model with maximal random structure failed to converge in that case ('degenerate Hessian with 1 negative eigenvalues'). We therefore report here the result of comparing models with maximal random structure for participants but only a random intercept for items.

[11] One could also look at the 6 tests (3 inferential tasks and 2 inference patterns) individually. They all converge to the same result: all of them show a significance difference between minimum and relative adjectives (all ps $< 4 \times 10^{-5}$), between color and relative adjectives (all ps $< 5 \times 10^{-6}$), and none does for the comparison between minimum and color adjectives (all p values between .04 and 0.93, and none of them passes the significance threshold after even minimal correction for multiple comparison).

through with color adjectives, which are not significantly distinguishable from minimum standard adjectives.

3.4 Discussion

First, this experiment confirms that the two entailment patterns clearly distinguish minimum standard absolute and relative adjectives. If therefore 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, and not significantly different than responses to minimum standard absolute adjectives.

The results of this experiment might be taken as strong support for the claim (advanced in Clapp 2012) that color adjectives are absolute adjectives with minimum standards. But one important limitation of this experiment is that it does not disambiguate the *quantitative* and *qualitative* readings of color adjectives.

Kennedy and McNally (2010) observe that an ambiguity 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 (21–22) demonstrate:

(21) The leaf is too yellowish to be perfectly green, but it is completely green.
(22) The leaf is only 90% green, but it is perfectly green.

Sentence (21) 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 (22) 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 (Fig. 2), which is associated with a scale of how much of an object is a particular color, and a *qualitative* reading (Fig. 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).[12]

Without a way of disambiguating these two readings, the fact that responses to color adjectives are significantly different than responses to relative adjectives in the entailment experiment might be due to the fact that, for example, the quantitative reading is minimum standard absolute while the qualitative reading is relative.

We conducted a second experiment that 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.

[12] An earlier observation of this distinction is made by Sapir (1944, p. 123): "Different examples of 'red' similarly exhibit 'mores' and 'lesses' with respect to intensity, size of surface or volume characterized as red, and degree of conformity to some expected standard of redness". See also McNally (2005). For naturally occurring examples of both readings, see Tribushinina (2008, p. 84).

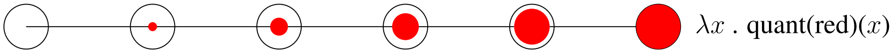


Fig. 2 Quantitative scale of redness

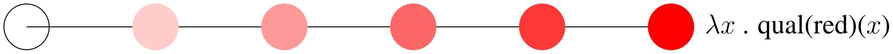


Fig. 3 Qualitative scale of redness

4 Experiment #2: The presupposition assessment task

Our second experiment employs versions of the tasks described in Syrett et al. (2010), which 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 Fig. 4a], and A's goal is to get B to pass over one of the rods. In such a context, A cannot felicitously use [(23)] 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 [(24)] 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 [(25)] to request the longer of the two rods.
>
> (23) # Please give me the red rod.
> (24) # Please give me the blue rod.
> (25) Please give me the long rod. (Syrett et al. 2010, p. 5).[13]

Syrett et al. (2010) found that participants 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, like "full" and "spotted".

Consider the pair of jars in Fig. 4b. 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 two incompletely full jars (handing neither or handing both counted as rejecting the request). 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.[14]

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

[14] 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

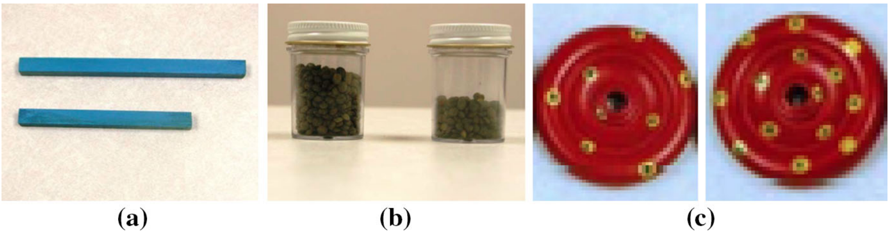


Fig. 4 Examples from Syrett (2007, Appendix E). **a** Please give me the long rod, **b** please give me the full one and **c** please give me the spotted one

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. (2010), 96% of adults rejected the request for "the spotted one" when both disks had some spots on them (as in Fig. 4c).

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 (2012, 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
>
> (26) 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 if it exceeds the minimal semantically encoded standard, which does not vary across contexts and cannot be adjusted through accommodation.

We found Clapp's intuition about the behavior of color adjectives surprising, and it did not accord with our 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. However, one of the authors did share

Footnote 14 continued
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"!

Clapp's judgment when it was the *quantitative* reading that was at issue. But these 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 our judgments 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* (where that means the relevant object is more of that color than not of that color). Adequately assessing these different predictions required 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 Participants

We recruited 42 participants over Amazon Mechanical Turk and paid them $0.80 each (see the Participants section in Experiment 1 for further details). 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.

4.2 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. 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. We asked subjects to respond to pictures of aliens with 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 Fig. 5).

4.2.1 Instructions and order of presentation

Participants were instructed in how to use the presupposition task by being given three clear examples: one example in which there is a clear correct response to the request to "click on the *Adj* alien", an example in which there is clear existence failure, requiring the "Neither is!" response, and an example in which there is clear uniqueness failure, requiring the "Both are!" response (see the Appendix for examples of the practice items). The order in which participants saw test items was randomized.

4.2.2 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

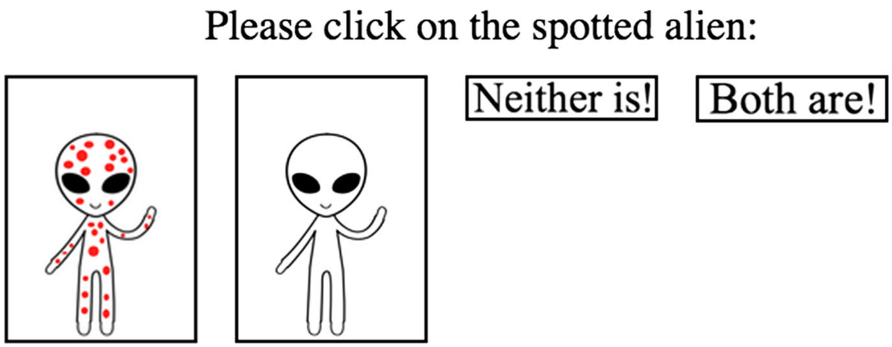


Fig. 5 Alien selection task

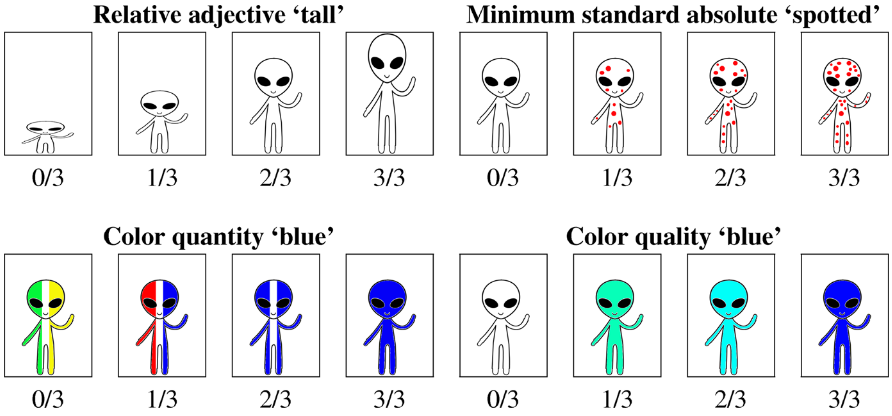


Fig. 6 Examples of aliens used to construct the different conditions for adjectives of each of the four relevant types

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").

4.2.3 Raw material: aliens with different degrees of adj-ness

The conditions were composed of two aliens. Individual aliens satisfied the adjectives to different degrees, as illustrated in Fig. 6, and discussed in detail in Appendix B.

4.2.4 Experimental conditions

Experimental conditions were obtained by juxtaposing 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 together with a 3/3 alien. We constructed the following experimental conditions for all adjectives:

Table 3 Predicted patterns of responses to the different conditions. The patterns are the same for all the control conditions, while different patterns in the target conditions identify different types of adjectives

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 standard		
NEITHER	CORRECT	CORRECT
relative + some threshold effect		
CORRECT	BOTH	BOTH
absolute + low standard		
NEITHER	CORRECT	BOTH
absolute + medium standard		
NEITHER	NEITHER	CORRECT
high standard		

- 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).

The total number of non-practice 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) × (3 control + 3 test) conditions = 72) × 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.

4.2.5 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

4.3 Results: Controls in the presupposition assessment task

Consider the three control conditions represented in Fig. 7: 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

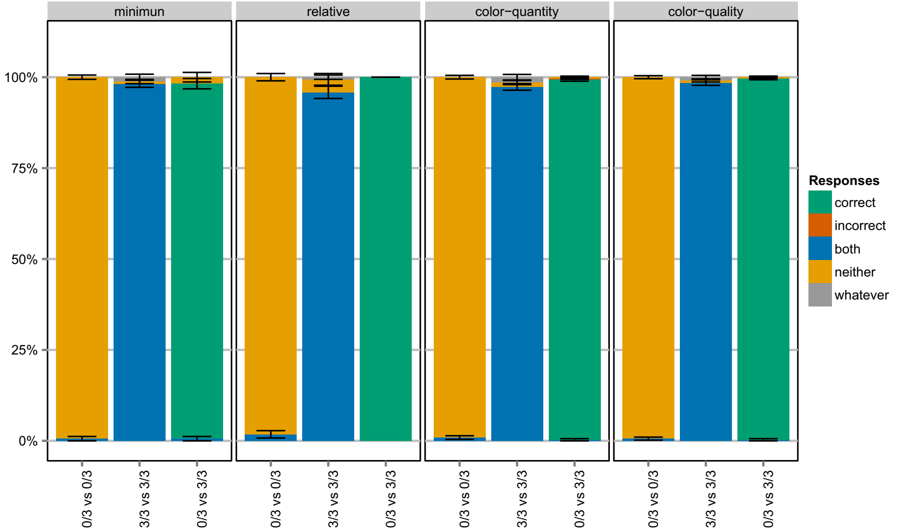


Fig. 7 Mean percentage of each response types for the control conditions in the presupposition assessment task

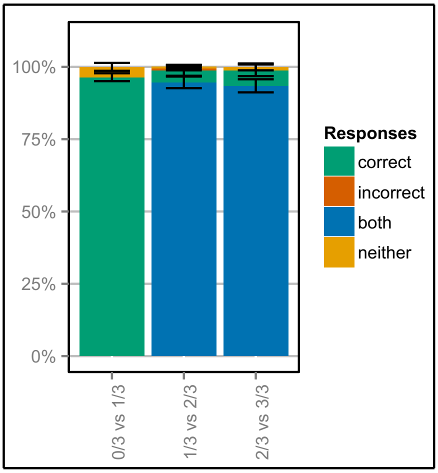
is evident from the stacked bar graph, participants are performing at or near ceiling with the control items (expected responses at least 95% of the time in each control condition for each type of adjective).

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 completely red aliens, with the response “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 and discussion: 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 Fig. 8, 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). In other words, when one of the aliens has *some degree* of



0/3 vs 1/3	1/3 vs 2/3	2/3 vs 3/3	Count
CORRECT	**BOTH**	**BOTH**	42
absolute + low standard			

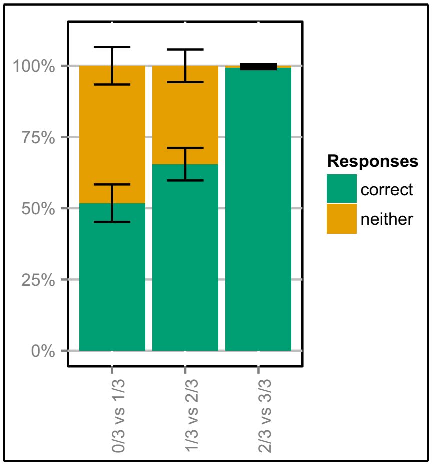
Fig. 8 Responses for minimum standard adjectives, (**a**) in the population, (**b**) counts of individuals displaying the corresponding pattern in the relevant row (where individuals are classified as having given a particular response when they give it at least 50% of the time), with an interpretation for the given pattern of responses (when possible)

the relevant property and the other does not (0/3 vs. 1/3 condition), then participants will respond to the request by clicking on the alien with *more* of the relevant property. But once both aliens have some degree of the relevant property (the 1/3 vs. 2/3 and 2/3 vs. 3/3 conditions), then participants will refuse the request by clicking on the "both are!" option. The table in Fig. 8 reveals that all participants (42) responded with that distinctive pattern.

Here and throughout our analysis of the presupposition assessment task, we assume that a participant shows a meaningful preference for a particular answer in a particular condition if she chose that answer at least 50% of the time across the repetitions (note that chance on each trial is at 25% given that there are 4 response choices). Adopting lower thresholds or using the majority response would yield the same results. With higher thresholds, however, many participants' profiles are undetermined because among the three test conditions, it is frequent that one fails to meet the stricter criterion.

Now consider the pattern of responses to relative adjectives represented in Fig. 9. 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 to binary values), all p values are below .005).[15] That

[15] Some conditions had very few repetitions (down to 4), which does not allow in general a proper, useful, or stable estimation of random effect for items. Throughout this experiment, then, we report analyses from comparisons of generalized mixed binomial models fitted to data after aggregation per participant, with maximal random structure for participant.



0/3 vs 1/3	1/3 vs 2/3	2/3 vs 3/3	count
CORRECT	CORRECT	CORRECT	29
relative standard			
NEITHER	CORRECT	CORRECT	7
relative + some threshold effect			
NEITHER	NEITHER	CORRECT	6
high standard			

Fig. 9 Responses for relative adjectives, see Fig. 8 for details of how to interpret the subject counts

confirms the findings in Syrett et al. (2010). Focusing on the 2/3 versus 3/3 condition (on the far right of the bar chart in Fig. 9), 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. If the standard view of accommodation with relative adjectives spelled out in Syrett et al. is correct, then participants should respond in the same way in the 0/3 versus 1/3 and 1/3 versus 2/3 conditions as well. That is, they should be willing to click on the alien with *more of* the relevant property in each condition:

> 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 (Syrett et al. 2010, p. 11).

But our results indicate that a significant number of subjects don't accommodate with relative adjectives that way. In the 0/3 versus 1/3 and 1/3 versus 2/3 conditions, there is some amount of "neither are!" responses (M = 52 and 35%, respectively), which should only characterize *maximum standard* adjectives on the standard view.

There are several ways that the surprising behavior of relative adjectives in these conditions might be reconciled with the standard picture. Syrett et al. (2010, p. 5 n. 3) 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 effect" shows up as the "inability or unwillingness on the part of speakers to …distinguish between objects that are very similar to each other relative to the scalar property that the predicate encodes". So, for example, if I ask you to "Click on the tall alien" when one alien is only slightly taller than the other, you might refuse. The threshold effect is due, according to Syrett et al., to the underlying vagueness of relative adjectives (the unwillingness to make crisp distinctions with relative adjectives also drives the sorites paradox). But if this "threshold effect" due to vagueness explains the failures to accommodate in the 0/3 versus 1/3 and 1/3 versus 2/3 conditions, then it should generate similar failures to accommodate in the 2/3 versus 3/3 condition, since the heights and widths 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 versus 1/3 condition than there is in the 1/3 versus 2/3 and 2/3 versus 3/3 conditions. That should make it *easier* to recognize a difference in height in the 0/3 versus 1/3 condition. But participants don't fail to accommodate in the 2/3 versus 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 versus 1/3 and 1/3 versus 2/3 conditions.[16]

A second, more plausible explanation of the surprising results that is still compatible with the standard picture of the meaning of relative adjectives involves a potential *order of presentation* effect in the experiment: Participants might not just be comparing the two aliens on the screen, but also the aliens they have seen in the experiment so far.[17] 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, and it would do so in a way consistent with the standard theory.[18]

One way to test this possibility is to see if this response (the "Neither is!" response to relative adjectives in the 0/3 versus 1/3 and 1/3 versus 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

[16] An anonymous referee wondered whether a Syrett et al. might respond by claiming, along the lines of Graff (2000) "interest-relative" account of vagueness, that what counts as "very similar" is interest (purpose and desire) relative, so if participants' interests shift between different contexts, two equal differences in magnitude (height, e.g.) may count as significant in one context and insignificant in another. We acknowledge that that is possible, but to make that response compelling, there would need to be some explanation of why participants' interests would shift across the experimental conditions we constructed. The next response we consider, which focuses on possible order of presentation effects, offers a more principled explanation of why responses may differ in the way we observed.

[17] Syrett et al. (2010, pp. 16–18) investigate a potential order of presentation effect that they thought might be affecting the responses of children in their experiment. They found that children were much more willing than adults to accommodate requests for "the full one" when presented with less than maximally full jars, and they wondered whether the difference might be a result of the children being primed by requests involving relative adjectives before the request for "the full one". They conducted an experiment that reversed the order of presentation, but found that the order of relative versus absolute adjectives had no significant effect on children's responses.

[18] Thanks to James Hampton and Robert van Rooij, who made this suggestion in discussion.

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 in the 0/3 versus 1/3 condition ($\chi^2(1) = 4.76$, $p = .029$). On the other hand, that result does not pass correction for more than one comparison, which is crucial given that there isn't evidence of such an increased rate of "neither" responses in the 1/3 versus 2/3 condition ($\chi^2(1) = 1.19$, $p = .28$). Moreover, there is also an effect on the rate of "both" responses to minimum standard adjectives in the 2/3 versus 3/3 condition as the experiment goes on ($\chi^2(1) = 6.88$, $p = .0087$). That suggests that participants are simply performing the task differently by the end of the experiment across different types of adjectives. Further experiments could be specifically designed to investigate these ordering effects more systematically.

We think the reactions to relative adjectives we observed is due to the presence of a lower *threshold* on the relevant scale, which objects have to cross before they are treated as having a shiftable standard.[19] Consider the possibility that what appear to be relative, minimum standard absolute, midpoint standard absolute and maximum standard absolute adjectives all share the same general, abstract type of standard, which is composed of three elements: a *lower threshold*, an *upper threshold*, and a *standard* (see Fig. 10).

To motivate the more abstract structure of thresholds and 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), but it is consistent with the failures to accommodate at the lower end of the scale that we observed in our second experiment.

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

[19] See Tribushinina (2008) for a discussion of the related notion of "cognitive reference points".

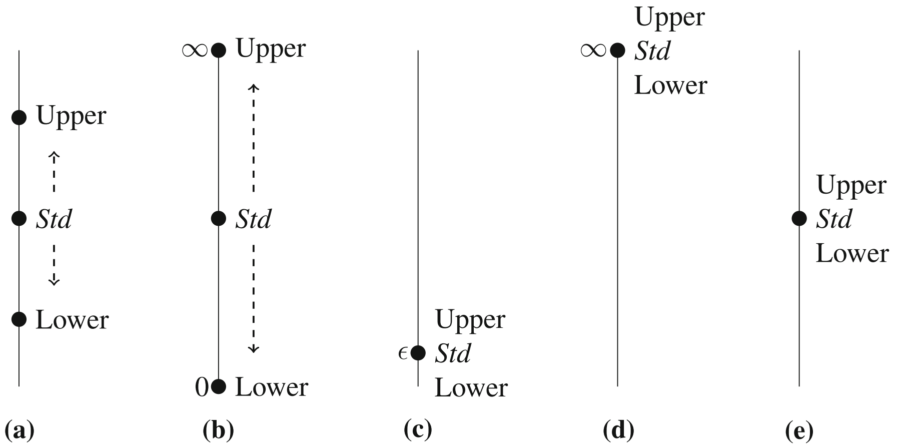


Fig. 10 A typology of adjectives based on a lower threshold (*lower*) and an upper threshold (*upper*), which together delimit the area of the scale where an adjective is relative, with a contextually variable standard in between (*Std*). The general, hybrid structure of adjectives is given in (**a**). Traditional relative adjectives are obtained when Upper and Lower are at the extremes of the scale (**b**). If Upper and Lower collapse 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)

by features of maximum standard, minimum standard, and relative adjectives could be described as having *hybrid* standards.

It is possible to derive all of the existing types of standard from this more abstract threshold and standard picture. 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 Fig. 10b). 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 Fig. 10c). A traditional maximum absolute standard is equivalent to collapsing the lower and upper thresholds at the maximum degree on the scale (see Fig. 10d). 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 Fig. 10e.

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), and also for the existence of hybrid standards.

An alternative (but closely related) explanation of the fact that some participants refuse to accommodate the existence presupposition associated with "Click on the tall alien" in the 0/3 versus 1/3 and 1/3 versus 2/3 conditions invokes the idea of a "geometrical constraint" on what can count as tall or wide: Only objects that have a sufficiently large ratio between their vertical and horizontal extents will be classified as tall or as wide (Barner and Snedeker 2008; Lang 1989).[20] That is, the aliens that are slightly taller than the shortest aliens are still too squat to count as tall, and the aliens that are slightly wider than the narrowest aliens are still too thin to count as wide. According to this proposal, in order to count as "the tall alien", an object would have to satisfy the relevant geometrical constraint (i.e. it can't be too squat).[21]

We find the existence of such a geometrical constraint on the application of the positive form of adjectives very plausible—indeed, it is a more specific way of understanding our idea of a lower threshold that objects have to cross before subjects will accommodate the existence presupposition. The geometrical constraint explanation would require introducing a second scale that measures ratios of height to width, rather than just degrees of height. That requires a more complex theoretical apparatus than the one we employ—we merely propose to identify a lower threshold on the same scale of height used to evaluate tallness, while the geometrical constraint proposal marks a similar type of threshold on *another* scale. In any event, these proposals are consistent with the refusals to accommodate at the lower end of the scale that we observed, and in that sense implement the same type of revision to the standard account of the meaning of scalar adjectives. For example, these proposals conflict with the prediction in Syrett et al. (2010, p. 11) that speakers should always be willing to accommodate the existence presupposition in requests featuring relative adjectives, like "click on the tall alien". Further experiments are needed to test for the existence of lower and upper thresholds and hybrid standards, and to determine whether it is such thresholds or a more specific geometrical constraint that best explains the unexpected behavior of relative adjectives that we observed.

4.5 Results and discussion: Color quantity

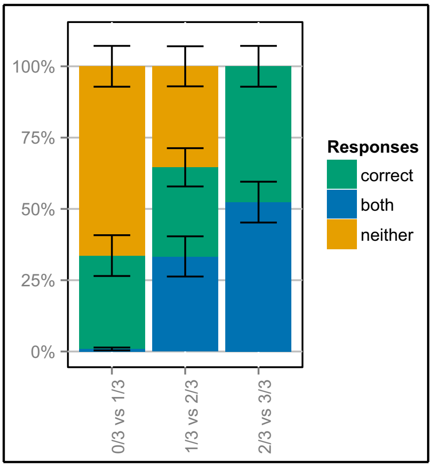
We now turn to consider how participants respond to the quantitative reading of color adjectives, represented in Fig. 11.

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.[22] What explains the difference? Do different participants respond to color quantity adjectives as if they were minimum standard and others respond to them as if they were relative? Are participants responding inconsis-

[20] Thanks to an anonymous referee for suggesting this explanation.

[21] A referee wondered whether participants might be using the box surrounding the aliens to set the lower threshold for "tall" and "wide"—in effect at the maximum height and width of the box, respectively. Such a heuristic might be guiding the responses of some participants, and it would explain the responses of those who gave a "high standard" response for the relative adjectives. But it wouldn't explain what is going on with those participants who seem to be operating with a midpoint standard, or those who behaved as predicted with relative adjectives. Ultimately, we don't know how participants pick the thresholds they operate with, and the "box heuristic" doesn't explain the overall pattern of results we observed.

[22] We can show this using our usual logit models. First, the proportion of NEITHER responses is higher for color than for minimum standard adjectives both in 0/3 versus 1/3 and in 1/3 versus 2/3 conditions ($p < .001$). Second, the proportion of BOTH responses is higher for color than for relative adjectives both in the 1/3 versus 2/3 condition and 2/3 versus 3/3 conditions ($p < .001$).



0/3 vs 1/3	1/3 vs 2/3	2/3 vs 3/3	count
CORRECT	BOTH	BOTH	13
absolute + low standard			
NEITHER	NEITHER	CORRECT	10
high standard			
NEITHER	CORRECT	BOTH	9
absolute + medium standard			
NEITHER	CORRECT	CORRECT	3
relative + some threshold effect			
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

Fig. 11 Responses for the quantitative readings of color adjectives. The question mark indicates a failure to choose a response consistently (at least half the time in the relevant condition)

tently? Do color quantity adjectives break the standard mold for classifying scalar adjectives?

By looking at individual responses in the table of Fig. 11, 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 standard (**CORRECT–BOTH–BOTH**) pattern, an absolute + high standard pattern (**NEITHER–NEITHER–CORRECT**), and an absolute + medium standard (**NEITHER–CORRECT–BOTH**) pattern. These three dominant patterns are followed by a motley tail of responses that don't clearly align with any standard.[23]

The absolute + medium standard pattern conforms with the account of the quantitative reading of color adjectives given in Kennedy and McNally (2010) and McNally (2011), in which for something to count as having a certain color, that color has to

[23] To assess which set of patterns are significantly populated by participants beyond chance, one can run successive χ^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 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. The last line of the table in this footnote reports the computation for that same fourth pattern but assuming that there are 34 unobserved patterns, and only with that many hidden alternatives in which responses could be diluted does the p value get below the .05 significance threshold, which is not sufficient if we take into account the need for correction for multiple comparisons.

"predominate", but once the color predominates, the color adjective behaves like an absolute adjective.[24]

The variation in where participants locate the standard indicates that while two 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) describe responses of different subgroups of participants, neither of those hypotheses fully captures the variety of how subjects respond to the quantitative reading of color adjectives. That is, the variety in standard types that exists in the theoretical literature mirrors the variety of responses that exist in the wider population that we sampled. This is a clear demonstration of the value of conducting a formal experimental investigation when the judgments of theorists conflict—it reveals that theorists who appear to have conflicting judgments may in fact be giving correct, but partial, descriptions of the linguistic facts.

One question raised by responses to the quantitative reading concerns entailments that color adjectives license. If the standard is located anywhere other than the minimal degree on the scale, then the entailments that characterize minimum standard adjectives, which were tested in our first experiment, would no longer hold. That includes McNally's intermediate absolute standard for the quantitative reading of color adjectives and the high (possibly maximum) standard, both of which we found evidence of in the presupposition accommodation experiment.

If, for example, 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?

Footnote 23 continued

Distribution		χ^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

[24] A referee pointed out that this pattern is also compatible with the following type of response to the prompt: the instruction "Click on the blue alien" could be read as "Click on the alien that is blue *and no other color*". This could be the result of a kind of scalar implicature, based on the potential competition between 'blue' and phrases such as 'blue and X' with 'X' denoting some other color. The competition between these readings would lead to effects embedded in the downward-entailing restrictor of the definite description, which itself is embedded in a non-truth conditional imperative sentence (therefore with no classic logical relation). Such a reading would thus require a particular set of assumptions about scalar alternatives and the mechanisms of Chierchia et al. (2012) to handle them. But on the assumption that such a reading exists, neither the 0/3 nor the 1/3 aliens satisfy this "is blue and no other color" predicate (the 0/3 alien has no blue, the 1/3 alien has another color besides blue), while the 2/3 and 3/3 aliens do satisfy the "is blue and no other color" predicate. Therefore, according to the proposed reading, we expect NEITHER, CORRECT and BOTH responses in the 0/3–1/3, 1/3–2/3, and 2/3–3/3 conditions, respectively. This potential analysis therefore predicts the pattern associated with the absolute + medium standard reading of color adjectives, but it does not predict the other response patterns that we obtained.

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.

4.6 Results and discussion: Color quality

Responses to the qualitative reading of color adjectives are represented in Fig. 12. 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.[25]

Looking at the counts of participants responding with different patterns, there is a majority absolute + low standard response pattern (**CORRECT–BOTH–BOTH**), followed by no clear pattern of responses.[26] The “?” response, which appears throughout the table in Fig. 12, indicates that the relevant participants did not choose any particular response at least half the time: more than a third of the participants (16 out of 42) responded in such a way.

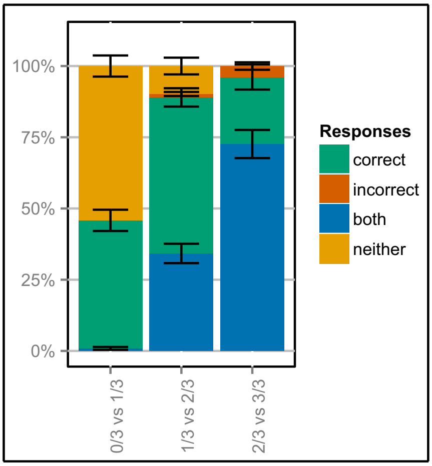
The noisiness of responses to the qualitative reading of color adjectives could be due either to non-semantic or semantic factors. The non-semantic factors could include interpersonal and intrapersonal variation in how participants perceive color or variation

[25] For instance, the qualitative reading of color adjectives generates more ‘neither’ responses than minimum standard adjectives in the 0/3 versus 1/3 condition ($p < 1 \times 10^{-14}$) and more ‘both’ responses than relative adjectives in the 1/3 versus 2/3 condition ($p < 1 \times 10^{-14}$).

One might wonder whether we chose the wrong examples of the 2/3 versus 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 Fig. 12), indicating that almost all participants agreed with our qualitative orderings.

[26] As argued in footnote 23, 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 unambiguous given that it is made of participants for which no clear response choice emerged in the 1/3 versus 2/3 condition.

Distribution		χ^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	0.28
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 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

Fig. 12 Responses for the qualitative readings of color adjectives, see Fig. 8 for details. Question marks indicate a failure to choose a response consistently (at least half the time in the relevant condition)

in the conditions in which the experiment was conducted (we couldn't control features of the screen or lighting conditions in which online participants perform the task).[27]

Another possible non-semantic explanation is that there is variation in participants' tolerance of *imprecision*. Syrett et al. (2010) found that children and adults behaved differently in response to the request "Please give me the full one" when confronted with two less than maximally full jars. Children (4 and 5 year olds) tended to accommodate the uniqueness presupposition and hand over the *fuller* of the two jars, while adults generally failed to accommodate. Assuming that "full" is a maximum standard absolute adjective that should block accommodation, the children's responses require some explanation. Syrett et al. explain the difference in terms of the children's "[greater] willingness to tolerate a certain amount of imprecision" (p. 18), where imprecision is a pragmatic phenomenon that permits the "use of a sentence or description that is false but 'close enough to true' for the purposes of the conversational exchange" (see Lasersohn 1999). Syrett et al. substantiate the claim that imprecision is at work in the different responses of the children in their study by showing that the children who accommodated with "full" had longer reaction times than they did when they accommodated with relative adjectives like "long", which they take to be evidence in support of the idea that there is an additional level of pragmatic processing taking place when children accommodate with "full". With this evidence in mind, it might be the case that the qualitative reading of color adjectives has a fixed absolute standard, but that there is variation among the adult population in how much imprecision is tolerated. Variation in tolerance of imprecision, plus subjective and situational perceptual varia-

[27] See Kuehni (2004) and Hansen (2015, 2016) for evidence and discussion of interpersonal variation in judgments about the "unique hues", and Morrison (2015) for reasons to suspect that there should be intrapersonal variation in color judgments.

tion could account for some of the variation on display in responses to the qualitative reading of color adjectives in the presupposition assessment task.[28]

Could variation in tolerance for imprecision also potentially account for the varied responses that we observed in response to the quantitative reading of color adjectives and the surprising refusals to accommodate with relative adjectives? It's not plausible that the patterns we observed with the quantitative reading of color adjectives can be accounted for this way, because we did not observe a greater willingness among some participants to accommodate—we observed *strict* responses that indicated participants were generally unwilling to accommodate, but that they differed in terms of where they located the relevant absolute standard on the scale. With regard to relative adjectives, it's not clear how differing tolerance for imprecision would be distinguishable from normal accommodation with the relative standard—and the fact that some participants were unwilling to accommodate at the lower end of the scale doesn't seem explicable as a case of different degrees of tolerance for imprecision in evaluation of the relevant standard.

One semantic explanation of the observed inter- and intrapersonal variation in response to the qualitative reading involves different types of semantic context sensitivity: There are semantic theories of scalar adjectives that interpret absolute adjectives as semantically context sensitive (van Rooij 2009; Toledo and Sassoon 2011). An alternative to the degree semantics for scalar adjectives discussed above is a semantics that treats scalar adjectives as partial functions from individuals to truth values, thereby partitioning the domain of individuals into a positive extension (individuals of which the adjective is "definitely true") , a negative extension (individuals of which the adjective is "definitely false") and an extension gap (individuals of which the adjective is neither definitely true nor definitely false) (Klein 1980). The domain of scalar adjectives is context dependent: contexts of utterance determine a comparison class of individuals that constitute the domain of the adjective. That allows for the clear context sensitivity of adjectives like "tall": if the domain of individuals determined by the context of utterance is Americans, then one of the authors of this paper will count as tall, but not if the domain of individuals are Oxford varsity rowers.

Van Rooij (2009), developing a version of this "delineation" approach to scalar adjectives, proposes that the domain of individuals relevant for the interpretation of absolute adjectives like "full" and "flat" is just the context insensitive "whole domain", so that an object is flat if and only if there is nothing flatter than it in the (whole) domain. But he makes room for a form of context sensitivity in absolute adjectives in terms of varying "standards of precision" (Lewis 1979). Roughly, van Rooij's proposal is that different (contextually variable) models can adopt different levels of precision in how individuals are ordered within the domain, in terms of the relations "Adj-er than" and "As adj as". So, for example, on a very coarsely grained model, Holland can fall within the positive extension of "flat" because the underlying measure structure doesn't distinguish between, e.g. Holland and Salar de Uyuni, "a salt flat in Bolivia that's the flattest place on earth", or, indeed, a perfectly flat geometrical plane.[29]

[28] Thanks to an anonymous referee for proposing the potential relevance of imprecision for explaining responses to the qualitative reading.

[29] http://www.cntraveler.com/stories/2014-07-28/salar-uyuni-maphead.

The idea of differing levels of precision might help explain the variation that we observed in responses to the qualitative reading of color adjectives. On first observing an object that is borderline blue (e.g. one of the aliens in the 1/3 qualitative condition), some participants might employ a coarsely-grained measure structure and place it in the positive extension of “blue”. Then, when they encounter a better example of blue (an alien in the 2/3 or 3/3 condition), they might be inclined to adopt a more precise measure structure that excludes the alien in the 1/3 qualitative condition from the positive extension of “blue”. (Or they might begin with a more exacting measure structure, then loosen their standards as the experiment goes on.) It’s conceivable that some participants are constantly and inconsistently adjusting the levels of precision that they think is appropriate for classifying the qualitative color of the aliens. That could be because, given the artificial context, and lack of prototypical anchors for classifying the aliens by color, they are simply unsure what the standards of precision are for what they are being asked to do.

Toledo and Sassoon (2011) also develop a semantics for scalar adjectives that interprets both relative and absolute adjectives in relation to a comparison class. Toledo and Sassoon propose that relative and absolute adjectives are interpreted relative to different *types* of comparison classes. The comparison class for absolute adjectives like “dirty” or “spotted” is comprised of the counterparts (“possible temporal stages”) of the individual the adjective is predicated of. The domain of individuals that will determine whether a surgical scalpel is dirty or not is determined by counterparts of the scalpel in “normal” possible worlds and times (counterparts of which would mostly be extremely clean, indeed sterilized), in contrast with the domain of individuals that will determine whether a particular kitchen knife is dirty (counterparts which would have a lower standard of overall cleanliness). In contrast, the comparison class for adjectives like “tall” or “wide” consists of “other members of the category containing the individual the adjective is predicated of, including distinct individuals in the index of evaluation” (p. 142). So the comparison class for “tall”, when it is predicated of a kindergartner, might include all other kindergartners in a particular school (for example). So, in short, whether a particular absolute adjective applies to an object depends on possible “normal” states of that object, while relative adjectives depend on the actual state of objects of the same relevant type.

The Toledo and Sassoon account could also help explain the pattern of interpersonal and intrapersonal variation we observed in response to the qualitative reading of color adjectives.[30] It’s not obvious which of the two types of comparison classes is appropriate for assessing color quality: whether someone’s face is red might be interpreted as a stage-level property of the face, and so be compared against counterparts of the same face, or it might be treated as an individual-level property and compared against objects of the same type. And those two interpretations could yield different classifications of an object in terms of its qualitative color: my face may turn red when I’m embarrassed—that is, red when compared with the normal state of my face—but even when it has turned red, it would not count as red when compared with how red some people’s faces are even in a neutral emotional context. Because of the lack of

[30] Thanks to an anonymous referee for proposing the relevance of the Toledo and Sassoon view for explaining the variation we observed.

clear context for assessing the target items in our experiment, participants' responses might be affected if they are ambivalent about which of these two types of comparison class is relevant for assessing the color of the aliens.

5 Philosophical consequences for the radical contextualist debate

How do our results bear on philosophical debates about the nature and extent of context sensitivity? Our results settle one important disputed issue in that debate, about the types of semantic (as opposed to pragmatic) resources that are available for explaining the context sensitivity that characterizes color adjectives. But our results also show that several other types of variation need to be taken into consideration in any assessment of that debate.

Scalar adjectives like "long" and "expensive" have played an important role in recent debates in philosophy over how best to understand the interaction of context and linguistic meaning. For example, 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. And Cappelen (2012) has recently proposed a contextualist account of "intuitive" on the grounds that it is a scalar adjective.[31]

More than any other type of scalar adjective, color adjectives have occupied a particularly central place in philosophical 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 (see Recanati (2004), §9.4, Searle 1978, 1980; Travis 1989, 1997, and 2006, Ch.4). Color adjectives seem like an ideal kind of expression to cite in support of radical contextualism. They appear to display forms of context sensitivity that are hard to characterize in a systematic way. But for that very reason they are also where the radical contextualist attack can be turned back most effectively: if semantic theorists can provide a satisfying account of the contextual variation displayed by color adjectives, where radical contextualist claims about untamable context sensitivity seem most plausible, then there is reason to believe radical contextualist claims about other expressions can be explained in a systematic way as well. And this is exactly what defenders of truth conditional semantic theories have done: they have argued that if the context sensitivity of color adjectives can be explained using the resources that already exist to handle the context sensitivity of relative scalar adjectives like "long" and "expensive" (plus ambiguity

[31] The next paragraph is adopted from the exposition of radical contextualism in Hansen (2011). See Davies (2014) for critical discussion of this way of understanding radical contextualism.

and other semantic concepts that can explain different types of variation), then a central category of examples that have been taken to lend strong support to radical contextualism actually demonstrate the explanatory power of truth conditional semantic theory rather than challenging it (Hansen 2011; Kennedy and McNally 2010).[32]

Recently, Clapp (2012) has argued that if color adjectives turn out to be absolute, rather than relative, then an important resource for explaining the contextual variability of the truth conditions of sentences containing color adjectives will not be available to the defenders of truth conditional semantic theory, namely the presence of a shifting, context dependent standard. That could potentially lend additional weight to radical contextualist arguments. The upshot of our experimental investigation comports to some degree with Clapp's argument that color adjectives are absolute: color adjectives figure in entailment patterns that are characteristic of minimum standard absolute adjectives, and not the patterns that are characteristic of relative adjectives. And in the presupposition accommodation task, responses to the quantitative reading of color adjectives do reveal patterns characteristic of absolute adjectives (though, importantly, there is interpersonal variation in where the standard is located on the scale), and there is a significant minority who respond to the qualitative reading in a way that patterns with minimum standard absolute adjectives.

Our experimental results are, therefore, consistent with Clapp's argument that color adjectives don't have a contextually variable (relative) standard, and therefore that there is one fewer semantic resource for explaining the truth conditional variability of sentences containing color adjectives than was proposed in Hansen (2011). But showing that color adjectives lack a context sensitive standard 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. This possibility comports with observations from cognitive semantics about the importance of a speaker's cognitive "vantage point" or "perspective" when evaluating color quality (Tribushinina 2008, p. 98). So, for example, what counts as focal red shifts depending on whether wine or blood is being evaluated. That indicates that what *scale* is associated with "red" can shift in different contexts, but it doesn't affect the investigation of where the relevant *standards* are located on the relevant scale. It's the latter issue that has been our concern in this paper. Other reasons for thinking that the scale itself is context sensitive is the evidence that there is interpersonal variation in where normal perceivers locate the prototypical, "unique", hues, and that perceptual judgments of color are sensitive to multifarious observation conditions: distance, angle, contrast, and so on (see Arnkil 2013 for a comprehensive review of the factors that influence color perception, and Hansen 2016 for a discussion of the subject relativity of color scales). And as discussed above, not all semantic theories license the move from the absoluteness of an

[32] Other theorists have invoked indexicality (Rothschild and Segal 2009; Szabó 2001) and polysemy (Vicente 2015) as alternative ways of responding to radical contextualism.

adjective to its semantic context insensitivity (van Rooij 2009; Toledo and Sassoon 2011, e.g.).

The philosophical significance of our experimental results therefore does not decisively favor one side or the other in the debate over the plausibility of radical contextualism versus truth conditional semantics. The larger philosophical significance of our investigation comes from responses to the quantitative reading of color adjectives in the presupposition assessment task. Those responses reveal that there is interpersonal variation in where the absolute standard is located on the quantitative scale: some participants treat it as minimum standard (in accordance with Clapp's judgment), some treat it as a midpoint ("predominating") standard (in accordance with Kennedy and McNally 2010; McNally 2011), and a third group treat it as having a very high (possibly maximum degree) standard. These responses reveal that various conflicting armchair judgments might each be accurate, but partial, expressions of the underlying semantic reality.

6 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.[33]

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 judgment), some treat it as having a very high standard (possibly maximum standard), and other participants treat it as somewhere in between (in accordance with Kennedy and McNally and McNally's judgments). 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 various plausible possibilities to investigate in further research: that the variation is due to non-semantic factors (perceptual factors, variability in tolerance for imprecision), or to semantic factors (variation in what comparison class is being used to interpret the relevant color adjective), or both.[34]

[33] Some interesting recent studies of whether aesthetic adjectives ("beautiful", "elegant") are relative or absolute have found results that roughly parallel some of our findings about color adjectives. Liao and Meskin (2015), using the presupposition assessment task, found that responses to aesthetic adjectives patterned in ways that differed from both relative and absolute adjectives. And Liao et al. (2016) argue that aesthetic adjectives behave like relative adjectives according to some diagnostics, and like absolute adjectives on others. See Phelan (2016) for a critical study of the Liao et al. results.

[34] As suggested by a referee, future research might also investigate how our data can be interpreted in light of recent work on definiteness, according to which definite descriptions presuppose uniqueness but not existence (Coppock and Beaver 2012, 2015).

We also found a surprising pattern of failures to accommodate with requests involving relative adjectives ("Click on the tall alien", e.g.). Again there are both non-semantic explanations (in terms of an order of presentation effect) and semantic explanations (in terms of the presence of a "geometric constraint" on what can count as tall, or the existence of a lower threshold) of this surprising pattern. Further experimental studies could provide additional evidence that could determine if one or both of these factors is responsible for the failure to accommodate that we observed.[35]

The significance of these experimental findings for the debate about radical contextualism is that there is good reason to think that color adjectives should not be assimilated to paradigmatic relative adjectives, and that the extensive contextual variability color adjectives display can't be explained (even in part) in terms of a context sensitive standard. But there are other sources of semantic context sensitivity that might account for the extensive variability that is a characteristic of color adjectives, and our results with the quantitative reading of color adjectives show that there is a great deal of interpersonal variation in how we understand color adjectives. The existence of that form of variability in how we understand color adjectives should inform debates about how context and meaning interact.

Acknowledgements Thanks to Jon Anzalone, Heather Burnett, Dan Lassiter, Shen-yi Liao, Louise McNally, Eliot Michaelson, Stephanie Solt, Kristen Syrett, audiences at the Semantic Content seminar at All Souls College, Oxford, the University of East Anglia, the Philosophy Society at the University of Reading, the discussion group at 4 Les Gauguins, the members of CCCOM, the Philosophie Expérimentale workshop at the Institut Jean-Nicod, the Pervasive Context workshop (Reading/Peking), the workshop on "Gradability, Scale Structure and Vagueness: Experimental Perspectives" hosted by the Spanish Research Council, the Semantics, Pragmatics and Philosophy of Language Workshop at Cambridge University, The New York Philosophy of Language Workshop, the "Meaning, Metaphor, and Maimonides" conference at the University of Chicago, and the anonymous referees for L&P 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 No. 313610 and was supported by ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC, AHRC research network grant AH/L015420/1, "Pervasive Context in Natural Language: The Problems And The Solutions", British Academy Grant SQ120050, "Quantitative Methods in Experimental Philosophy of Language", and an External Faculty Fellowship at Stanford University's Humanities Center.

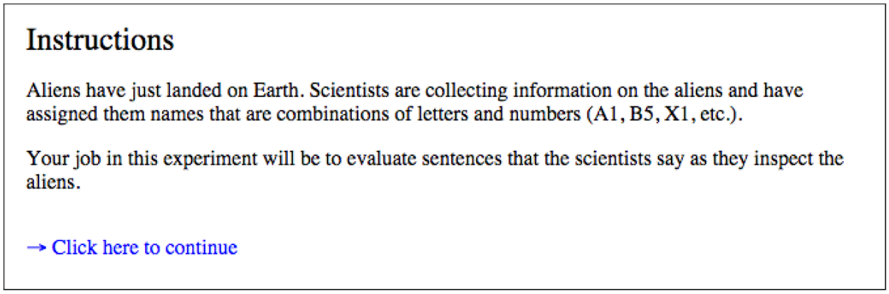
Appendix A: Instructions and practice items for experiments 1 and 2

Experiment 1

Participants in Experiment 1 were randomly assigned to one of two configurations in experiment 1: A "downward arrow"-first and "therefore/but"-second configuration, or

[35] To facilitate further studies, we have made our raw data, our R scripts, and our experimental stimuli available online at http://semanticsarchive.net/Archive/TVmMjMwY/Hansen-Chemla-ColorAdj.html.

a “therefore/but”-first and “downward arrow”-second configuration. In both configurations, participants saw the following generic instructions:

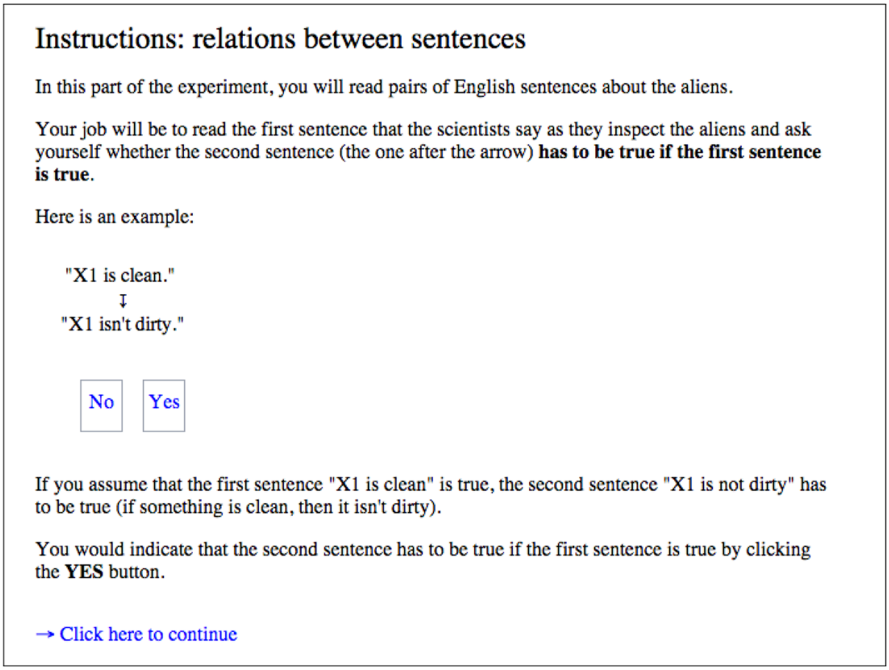


Instructions

Aliens have just landed on Earth. Scientists are collecting information on the aliens and have assigned them names that are combinations of letters and numbers (A1, B5, X1, etc.).

Your job in this experiment will be to evaluate sentences that the scientists say as they inspect the aliens.

Instructions for the “downward arrow” task read as follows:



Instructions: relations between sentences

In this part of the experiment, you will read pairs of English sentences about the aliens.

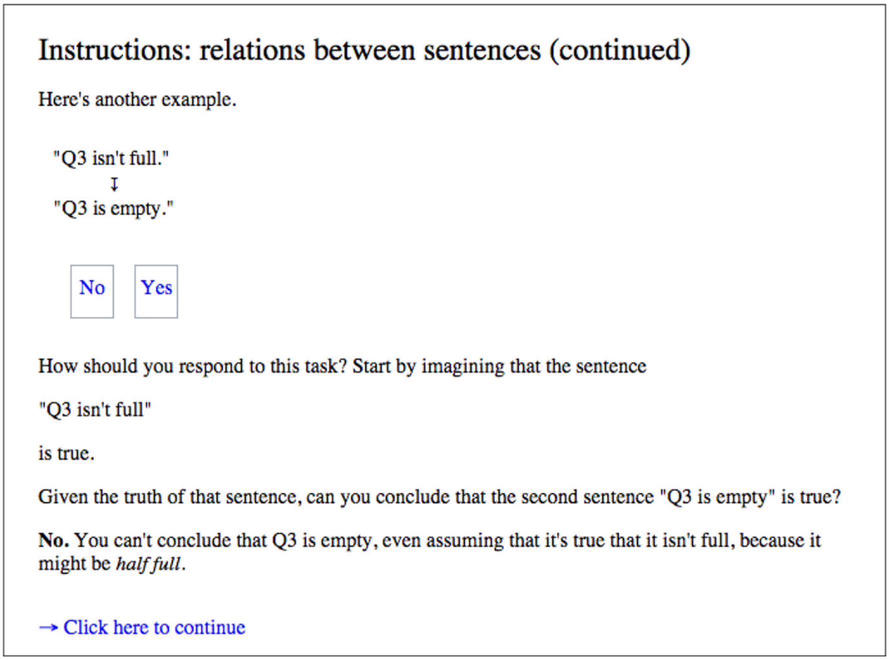
Your job will be to read the first sentence that the scientists say as they inspect the aliens and ask yourself whether the second sentence (the one after the arrow) **has to be true if the first sentence is true**.

Here is an example:

"X1 is clean."
↧
"X1 isn't dirty."

If you assume that the first sentence "X1 is clean" is true, the second sentence "X1 is not dirty" has to be true (if something is clean, then it isn't dirty).

You would indicate that the second sentence has to be true if the first sentence is true by clicking the **YES** button.



Instructions: relations between sentences (continued)

Here's another example.

"Q3 isn't full."
↧
"Q3 is empty."

How should you respond to this task? Start by imagining that the sentence

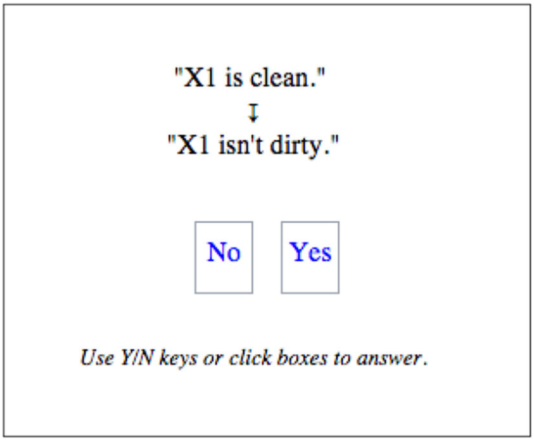
"Q3 isn't full"

is true.

Given the truth of that sentence, can you conclude that the second sentence "Q3 is empty" is true?

No. You can't conclude that Q3 is empty, even assuming that it's true that it isn't full, because it might be *half full*.

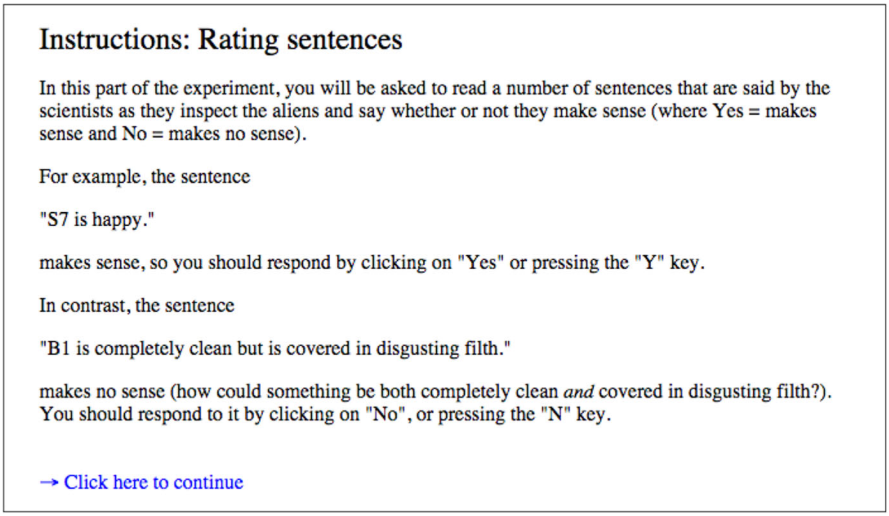
The instructions were then followed by one practice item, which repeated one of the examples from the instructions to give participants a chance to get used to the interface:



"X1 is clean."
↧
"X1 isn't dirty."

Use Y/N keys or click boxes to answer.

Instructions for the “therefore/but” task read as follows:



Instructions: Rating sentences

In this part of the experiment, you will be asked to read a number of sentences that are said by the scientists as they inspect the aliens and say whether or not they make sense (where Yes = makes sense and No = makes no sense).

For example, the sentence

"S7 is happy."

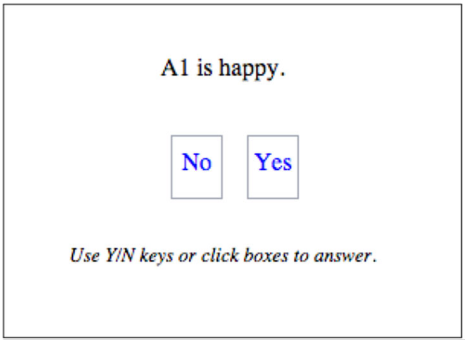
makes sense, so you should respond by clicking on "Yes" or pressing the "Y" key.

In contrast, the sentence

"B1 is completely clean but is covered in disgusting filth."

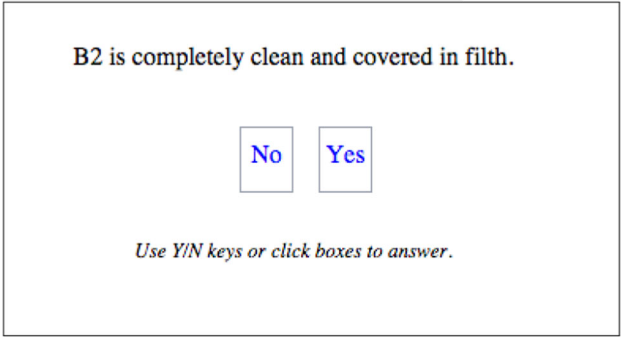
makes no sense (how could something be both completely clean *and* covered in disgusting filth?). You should respond to it by clicking on "No", or pressing the "N" key.

The instructions were then followed by two practice items, which repeated the two examples from the instructions to give participants a chance to get used to the interface:



A1 is happy.

Use Y/N keys or click boxes to answer.

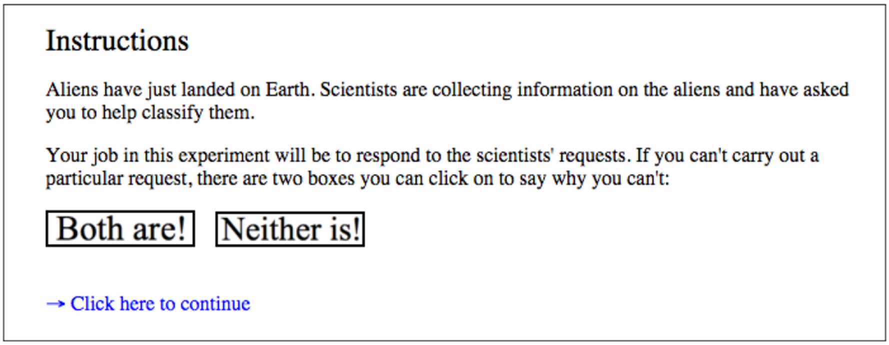


B2 is completely clean and covered in filth.

Use Y/N keys or click boxes to answer.

Experiment 2

Instructions for Experiment 2 read as follows:

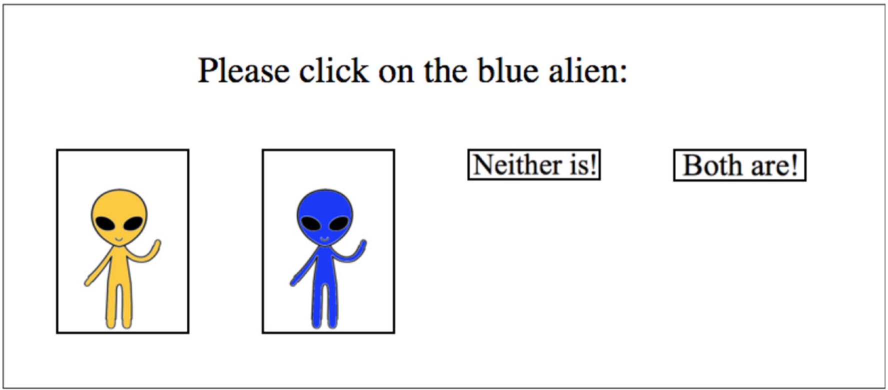


Instructions

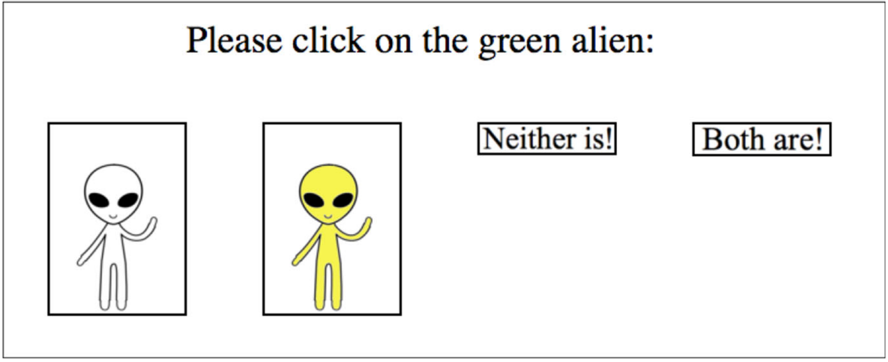
Aliens have just landed on Earth. Scientists are collecting information on the aliens and have asked you to help classify them.

Your job in this experiment will be to respond to the scientists' requests. If you can't carry out a particular request, there are two boxes you can click on to say why you can't:

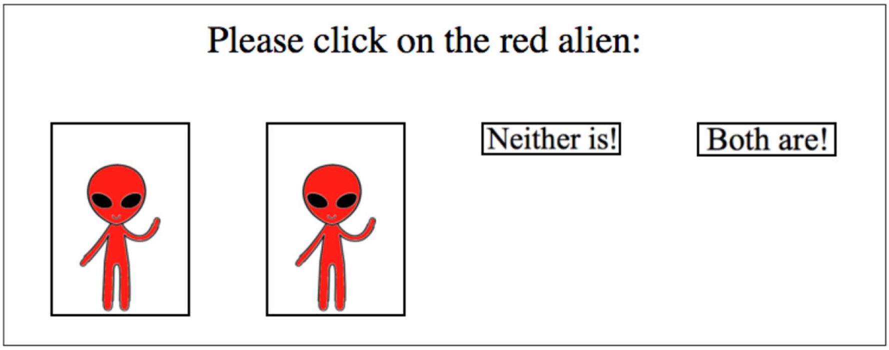
Participants then saw three practice items that familiarized them with *clearly correct* answers to requests, cases of *clear existence failure*, and cases of *clear uniqueness failure*:



Please click on the blue alien:



Please click on the green alien:



Please click on the red alien:

Appendix B: Methods for constructing stimuli in the presupposition assessment task

- 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 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.[36] 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). 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

[36] A referee points out that there is a potential artifact created by the fact that while the relevant color varies systematically, the other colors on the alien in the 1/3 condition also varied—so, for example, the 1/3 red alien has red, blue, and white stripes, while the 1/3 green alien has green, yellow, and white stripes. This could potentially affect results involving the 1/3 condition, but we don't think there is reason to expect this to bias results in any particular direction.

had 10 (mutatis mutandis for "dirty", which was generated by using clicks of the "spray can" function with grayish brown "dirt").

- 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 Fig. 6).

References

Arnkil, H. (2013). *Colours in the visual world*. Helsinki: Aalto Art Books.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-Year-olds interpret Tall and Short based on the size distributions of novel noun referents. *Child Development*, *79*(3), 594–608.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

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

Burnett, H. (2014). A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy*, *37*(1), 1–39.

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

Chierchia, G., Spector, B., & Fox, D. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics. An international handbook of natural language meaning* (Vol. 3, pp. 2297–2332). Berlin: Mouton de Gruyter.

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

Coppock, E., & Beaver, D. (2012). Weak uniqueness: The only difference between definiteness and indefinites. In A. Chereches (Ed.), *Proceedings of Semantics and Linguistic Theory (SALT)* (pp. 197–217). Ithaca, NY: CLC Publications.

Coppock, E., & Beaver, D. (2015). Definiteness and determinacy. *Linguistics and Philosophy*, *38*(5), 377–435.

Davies, A. (2014). Off-target responses to occasion-sensitivity. *Dialectica*, *68*(4), 499–523.

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.

Graff, D. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, *28*(1), 45–81.

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

Hansen, N. (2015). A new argument from interpersonal variation to subjectivism about color: A response to gómez-torrente. *Noûs*, (EarlyView):1–8.

Hansen, N. (2016). Color comparisons and interpersonal variation. *Review of Philosophy and Psychology*, online first:1–18.

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

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

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

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45.

Kuehni, R. (2004). Variability in unique hue selection: A surprising phenomenon. *Color Research and Application*, *29*, 158–162.

Lang, E. (1989). The semantics of dimensional designation of spatial objects. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation* (pp. 263–417). New York, NY: Springer.

Lasersohn, P. (1999). Pragmatic halos. *Language*, *75*(3), 522–551.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*(1), 339–359.

Liao, S.-Y., McNally, L., & Meskin, A. (2016). Aesthetic adjectives lack uniform behavior. *Inquiry*, *59*(6), 618–631.

Liao, S.-Y. & Meskin, A. (2015). Aesthetic adjectives: Experimental semantics and context-sensitivity. *Philosophy and Phenomenological Research*, Early View:1–28.

McNally, L. (2005). Lexical representation and modification within the noun phrase. *Recherches Linguistiques de Vincennes*, *34*, 191–206.

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *ViC 2009: Papers from the ESSLLI 2009 workshop on vagueness in communication* (Vol. 6517, pp. 151–168)., Lecture notes in computer science Heidelberg: Springer.

Morrison, J. (2015). Anti-atomism about color representation. *Noûs*, *49*(1), 94–122.

Panzeri, F., & Foppolo, F. (2012). Can children tell us something about the semantics of adjectives? In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, & M. Westera (Eds.), *Logic, language and meaning: 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19–21, 2011, revised selected papers* (pp. 170–179). Berlin: Springer.

Phelan, M. (2016). Gradability and multidimensionality in aesthetic adjectives. Unpublished ms.

Recanati, F. (2004). *Literal meaning*. Cambridge: Cambridge University Press.

Rett, J. (2015). *The semantics of evaluativity*. Oxford: Oxford University Press.

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

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

Sapir, E. (1944). Grading: A study in semantics. *Philosophy of Science*, *11*(2), 93–116.

Sassoon, G. W. (2012). A typology of multidimensional adjectives. *Journal of Semantics*, *30*(3), 1–46.

Searle, J. R. (1978). Literal meaning. *Erkenntnis*, *13*(1), 207–224.

Searle, J. R. (1980). The background of meaning. In F. Kiefer & M. Bierwisch (Eds.), *Speech act theory and pragmatics* (pp. 221–232). Dordrecht, Holland: Reidel.

Solt, S. (2016). Multidimensionality, subjectivity and scales: Experimental evidence. Unpublished ms.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

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

Syrett, K. (2007). *Learning about the structure of scales: Adverbial modification and the acquisition of the semantics of gradable adjectives*. Ph.D. thesis, Northwestern University, Evanston, IL.

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

Szabó, Z. G. (2001). Adjectives in context. In I. Kenesei & R. M. Harnish (Eds.), *Perspectives on semantics, pragmatics, and discourse: A Festschrift for Ferenc Kiefer* (pp. 119–146). Amsterdam: John Benjamins Publishing Company.

Team, R. C. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Travis, C. (1989). *The uses of sense: Wittgenstein's philosophy of language*. Oxford: Oxford University Press.

Travis, C. (1997). Pragmatics. In B. Hale & C. Wright (Eds.), *A companion to the philosophy of language* (pp. 87–107). Oxford: Blackwell.

Travis, C. (2006). *Thought's footing*. Oxford: Oxford University Press.

Tribushinina, E. (2008). *Cognitive reference points: Semantics beyond the prototypes in adjectives of space and colour*. Ph.D. thesis, Universiteit Leiden, Leiden, The Netherlands.

Unger, P. (1975). *Ignorance: A case for skepticism*. Oxford: Oxford University Press.

van Rooij, R. (2009). Vagueness in linguistics. In G. Ronzitti (Ed.), *The Vagueness handbook* (pp. 123–170). Berlin: Springer.

Vicente, A. (2015). The green leaves and the expert: Polysemy and truth conditional variability. *Lingua*, *157*, 54–65.

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