Color Adjectives, Standards and Thresholds: An Experimental Investigation

Nat Hansen
Department of Philosophy, University of Reading

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

January 7, 2015

Abstract

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

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

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

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

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 just in case the hardcover is mapped to a greater degree on the scale of costthan the paperback:

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

(8) $[\![\textbf{more expensive than}]\!]_{<<e;<e;t>>} = y\ x . \text{expensive}(x) \quad \text{expensive}(y)$

(9) $[\![\textbf{The hardcover is more expensive than the paperback}]\!]_t$ = expensive(the hard-cover) 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 un-pronounced ("null") morpheme that supplies the relevant comparison. So, for the purposes of semantic interpretation, a "bare positive" construction like (10) is actually understood as (11):

(11) The hardcover is posexpensive.

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

(12) $[\![\textbf{pos}]\!]_{<<e;d>;<e;t>>} = G\ x . G(x) \quad \text{standard}(G)$

(13) $[\![\textbf{pos expensive}]\!]_{<e;t>} = x . \text{expensive}(x) \quad \text{standard}(\text{expensive})$

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

So (14) is true just in case the hardcover is mapped to a degree 9038(9hat)-339(ctanda)-319(mut'224)-TJ 0

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

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

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

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

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

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

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

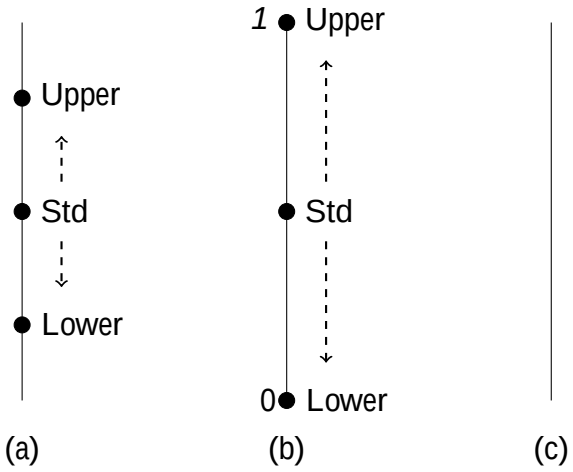
When a scalar adjective combines with "pos", whether the adjective is relative or minimum or maximum standard absolute is part of the input to the context sensitive "standard" function that is part of the meaning of "pos", which determines the adjective's standard value. With minimal and maximal standard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

1.2 Hybrid standards

The standard, "off the shelf" semantics for scalar adjectives described in the previous section provides a neat taxonomy olutetion pro

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

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



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

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

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

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

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

1.3 Color adjectives: Background

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

compositional truth conditional semantic theory (see Travis 2008b, Kennedy and McNally

Figure 2: Quantitative scale of redness

x . quant(red)(x)

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

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

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

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

> Interpretive Economy
>
> Maximize the contribution of the conventional meanings of the elements of a

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

3. So color adjectives are absolute.

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

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

3.1 Materials, design and task

3.1.1 Adjectives

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

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

Table 1: Target adjectives in the entailment experiment

3.1.2 Three inferential tasks

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

> The downward arrow task ("#") is intended to elicit more or less direct judgments about entailment. After a brief introduction to entailment, participants were asked to say whether a sentence following the downward arrow has to be true if the sentence preceding the arrow is true, as in (26a).
>
> The THEREFORE task is a linguistic translation of the "#" test: participants were asked to say whether a sentence of the form "p therefore q" makes sense, with p and q the appropriate premise and conclusion as in (26b).
>
> The BUT task was an anti-inference test, in which participants were asked to say whether a sentence of the form "p, but not q" makes sense, see (26c); negative responses here indicates entailment from p to q.

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

a. Downward arrow task "#":

"X1 is taller than Y2."

#

"X1 is tall."

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

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

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

3.1.3 Order of presentation

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

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

3.1.4 Predictions

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

	relative	minimum	color
"#"	no	yes	?
THEREFORE	no	yes	?
BUT	yes	no	?

Table 2: Predictions for the inferential tasks

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

3.2 Participants

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

3.3 Results of the entailment pattern experiment

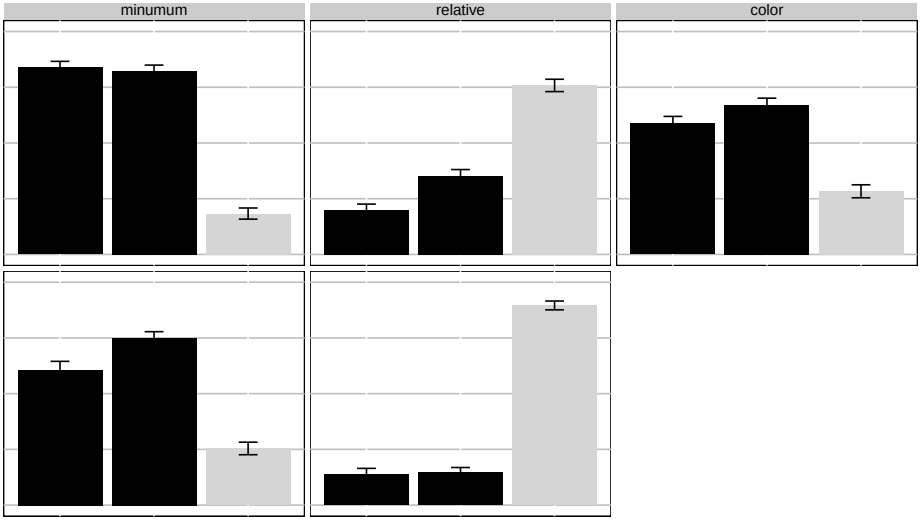


Figure 5: Percentages of "Yes" responses to the entailment pattern experiment

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

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

difference) are obtained for entailment pattern 2 (:042 for the “#” and :11 for BUT

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

(27) # Please give me the red rod.

(28) # Please give me the blue rod.

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

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

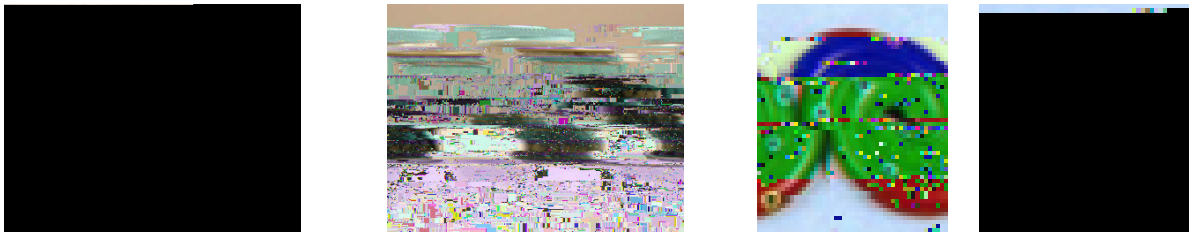


Figure 6: Examples from Syrett (2007, Appendix E)

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

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

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

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

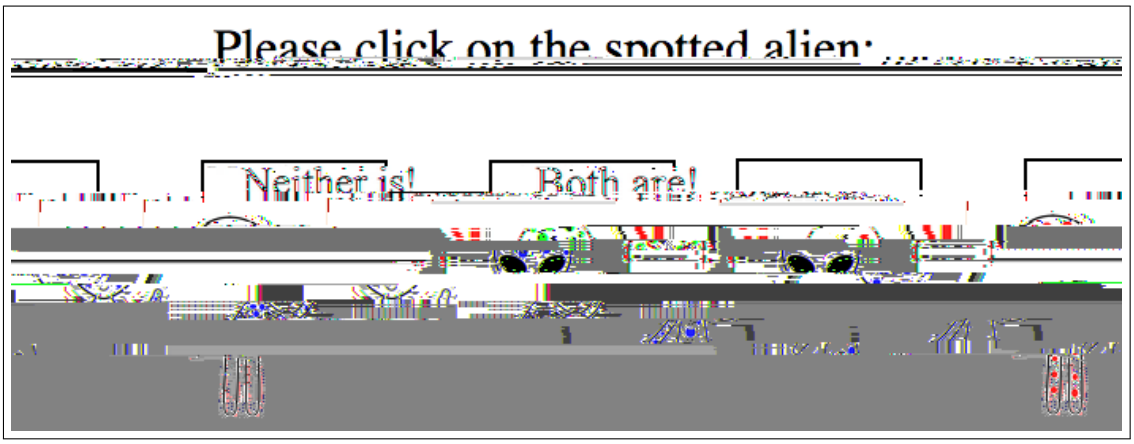


Figure 7: Alien selection task

4.1.1 Adjectives and adjective types

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

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

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

> Each adjective in the experiment is associated with a scale. A maximalcondition 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

Relative adjective ‘tall’

4.1.4 Response coding

Responses to the task were coded as follows:

CORRECT:	Clicking on the alien with moreof the relevant property
INCORRECT:	Clicking on the alien with lessof the relevant property
WHATEVER:	Clicking on either of the aliens when they are identical
NEITHER:	Clicking on the “neither” button
BOTH:	Clicking on the “both” button

4.2 Participants

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

4.3 Results: Controls in the presupposition assessment task

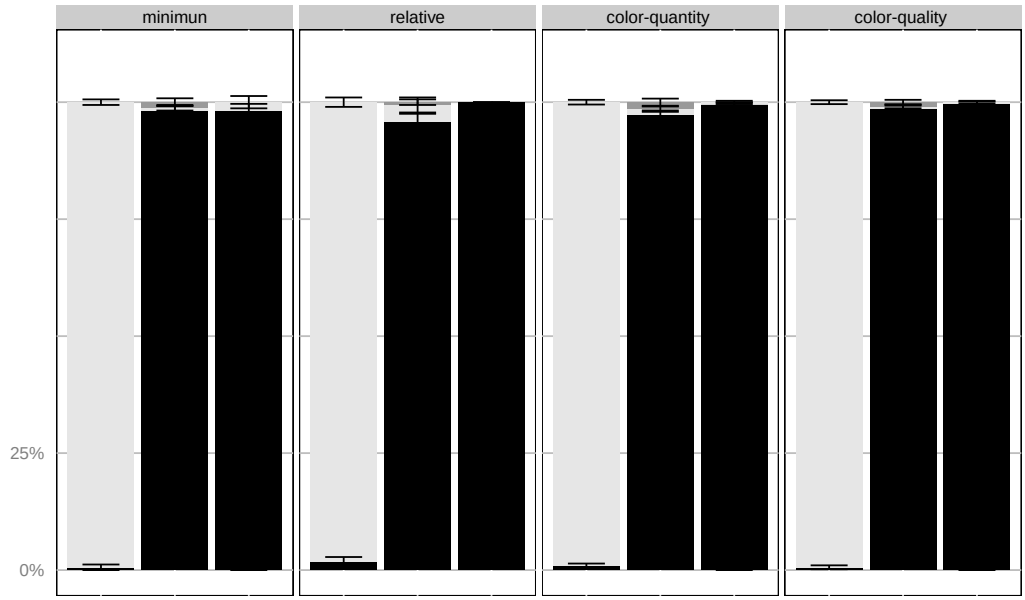


Figure 9: Mean percentage of each response types for the control conditions in the presupposition assessment task

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

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

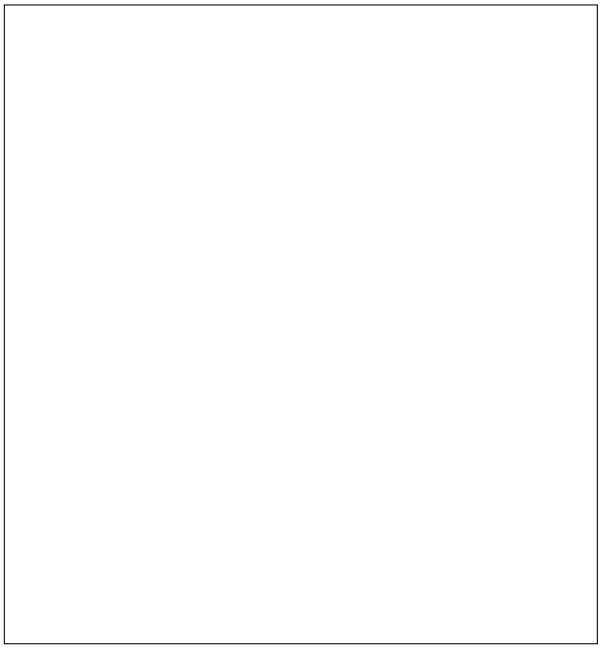
In the 0/3 vs 0/3 condition, participants almost universally responded with the response "neither are!", indicating existence failure. In the 3/3 vs 3/3 condition participants responded to the request to, e.g., click on the red alienwhen confronted with two equally, completely red aliens by responding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

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

4.4 Results: Minimum standard and relative adjectives

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

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



0/3 vs 1/3	1/3 vs 2/3	2/3 vs 3/3	Count
CORRECT	**BOTH**	**BOTH**	42
absolute + low threshold (and standard)			

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

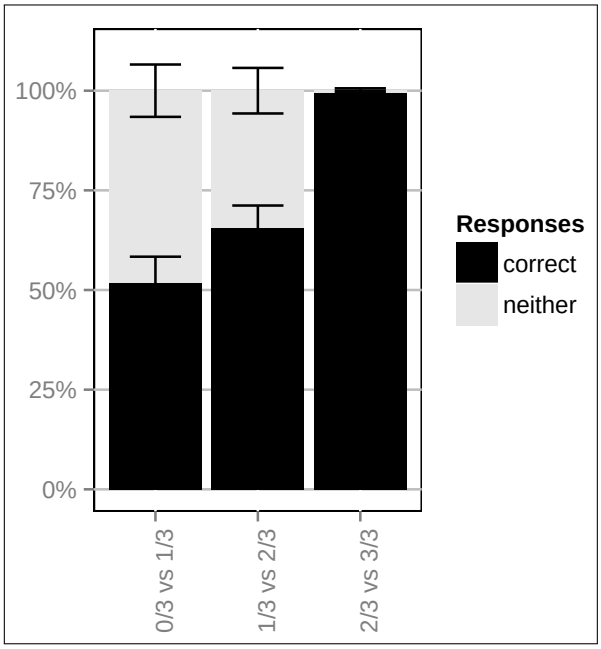
tives on the standard view. Syrett et al. (2010, p. 11) predict that participants will always be able to accommodate the uniqueness and existence presuppositions of definite descriptions when combined with relative adjectives (as in "Please click on the tall alien"):

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

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

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

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



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

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

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

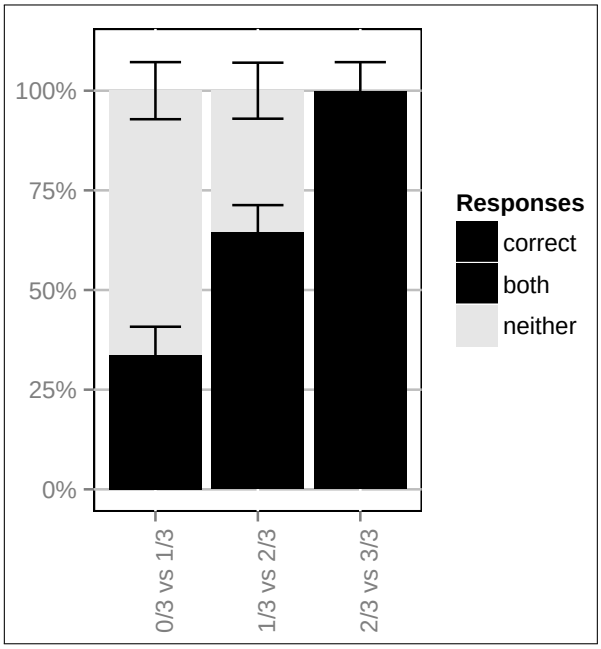
4.5 Results: Color Quantity

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

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

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

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



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

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

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

By looking at individual responses in the table of Figure 12, we get a more fine-grained picture of how the quantitative reading of color adjectives relates to relative and minimum standard adjectives. Responses to the quantitative reading of color adjectives fall mainly into three patterns: either an absolute + low threshold and standard (**CORRECT-BOTH-**

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

4.6 Results: Color quality

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

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

---

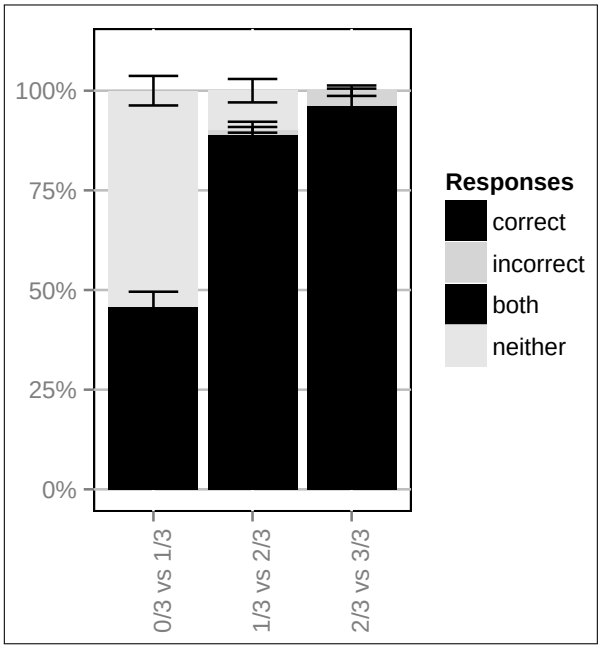
that we expect lower extremes), these tests tell us at each stage if the maximal extreme value that remains in the set does contribute a significant divergence from chance. The results are as in the table below, showing that the first three extreme patterns, with 13, 10 and 9 participants respectively,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

distribution		^{2}	p
13,10,9,3,1,1,1,1,1,1,1		54	5:10 ^{8}
10,9,3,1,1,1,1,1,1,1		39	1:10 ^{5}
9,3,1,1,1,1,1,1,1		27	7:10 ^{4}
3,1,1,1,1,1,1,1		2.8	:90
3,1,1,1,1,1,1,1	adding 34 empty cells (with 0s)	41	:048

[21] For instance, color adjectives generate more 'neither' responses than minimum standard adjectives in the 0/3 vs 1/3 condition ($p < 1:10^{14}$) and more 'both' responses than relative adjectives in the 1/3 vs 2/3 condition ($p < 1:10^{14}$).

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

distribution		^{2}	p
22,6,3,2,2,2,1,1,1,1,1		101	2:10 ^{16}
6,3,2,2,2,1,1,1,1,1		11	:28
6,3,2,2,2,1,1,1,1,1	adding 6 empty cells (with 0s)	15	:013



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

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

4.7 Discussion

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

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

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

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

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

4.7.1 Possible concerns

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

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

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

5 Conclusions and further research

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

Furthermore, the advantage of investigating the nature of standards for different types of adjectives using a formal experimental approach is that it reveals that various existing accounts of the standards appropriate for color adjectives are all only partially correct. It turns out that the quantitative reading involves interpersonal variation about where the standard is located: some participants treat the quantitative reading as minimum standard-like (in alignment with Clapp's prediction), some treat it as having a very high threshold (or possibly as maximum standard absolute), and other participants treat it as somewhere in between (in accordance with McNally's prediction). The qualitative reading, on the other hand, displays no clear pattern of responses beyond a majority minimum standard response. The explanation n lly's p is thatmi(8(')55(s)oachl9Nanchorusi1ly)-uusi1lyofiJ 0 gree17(is)-r-316

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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