Scale structure: Processing Minimum Standard and Maximum Standard Scalar Adjectives

Lyn Frazier, Charles Clifton, Jr. and Britta Stolterfoht

University of Massachusetts Amherst

Address correspondence to:

Lyn Frazier

Department of Linguistics

University of Massachusetts

Amherst, MA 01003 USA

lyn@linguist.umass.edu

Abstract

Gradable adjectives denote a function that takes an object and returns a measure of the degree to which the object possesses some gradable property (Kennedy, 1999). Scales, ordered sets of degrees, have begun to be studied systematically in semantics (Kennedy, to appear, Kennedy & McNally, 2005, Rotstein & Winter, 2004). We report four experiments designed to investigate the processing of absolute adjectives with a maximum standard (e.g., clean) and their minimum standard antonyms (dirty). The central hypothesis is that the denotation of an absolute adjective introduces a 'standard value' on a scale as part of the normal comprehension of a sentence containing the adjective (the "Obligatory Scale" hypothesis). In line with the predictions of Kennedy and McNally (2005) and Rotstein and Winter (2004), maximum standard adjectives and minimum standard adjectives systematically differ from each other when they are combined with minimizing modifiers like slightly, as indicated by speeded acceptability judgments. An eye movement recording study shows that, as predicted by the Obligatory Scale hypothesis, the penalty due to combining slightly with a maximum standard adjective can be observed during the processing of the sentence; the penalty is not the result of some after-the-fact inferencing mechanism. Further, a type of 'quantificational variability effect' may be observed when a quantificational adverb (mostly) is combined with a minimum standard adjective in sentences like The dishes are mostly dirty, which may receive either a degree interpretation (e.g. 80% dirty) or a quantity interpretation (e.g., 80% of the dishes are dirty). The quantificational variability results provide suggestive support for the Obligatory Scale hypothesis by showing that the standard of a scalar adjective influences the preferred interpretation of other constituents in the sentence.

Scale structure: Processing Minimum Standard and Maximum Standard Scalar Adjectives

Gradable adjectives like wet or straight or dirty have been analyzed as denoting a function (of semantic type <e,d>) that takes an object and returns a measure of the degree to which the object possesses some gradable property (e.g., the property of being clean; Kennedy, 1999, 2006, Kennedy & McNally, 2005). Rather than considering relative adjectives like tall or expensive, whose evaluation depends on the identity of the noun that they modify, we will focus here on absolute adjectives, which do not readily admit borderline cases, for example.[1] We will also restrict attention to absolute gradable adjectives that have antonyms such as clean-dirty, healthy-sick and dead-alive.

Kennedy and McNally (2005) draw a distinction between adjectives with a maximum standard (clean, dry, smooth, ...) and those with a minimum standard (dirty, wet, rough , ...) (for earlier work, see Cruse, 1980, Kamp & Rossdeutscher, 1994, Yoon, 1996). Scales (an ordered set of degrees corresponding to the extent to which an object exhibits a gradable property) have begun to be studied systematically in linguistics (see in particular Kennedy, to appear, Kennedy & McNally, 1999, 2002, Rotstein & Winter, 2004). In the case of absolute adjectives, the basic

[1] Thus, they do not give rise to the Sorities Paradox (Kennedy, to appear.) characteristic of vague predicates, e.g., relative adjectives like expensive:

(i) (=Kennedy's (2)) The Sorities Paradox

P1. Any $5 cup of coffee is expensive (for a cup of coffee).

P2. Any cup of coffee that costs 1 cent less than an expensive one is expensive (for a cup of coffee.)

P3. Therefore, any free cup of coffee is expensive.

idea is that some have as their default value an interpretation determined by the maximum degree on their scale. These are termed 'maximum standard' adjectives. Clean is an example. Something is clean only if it is free of dirt. Maximum standard adjectives contrast with other (frequently complementary) adjectives whose default interpretation only requires that an entity exhibit the property denoted by the adjective to some non-zero degree. These are termed 'minimum standard' adjectives. Dirty is an example. Something is dirty if it has some non-zero amount of dirt.

Maximum standard and minimum standard adjectives are claimed to give rise to distinct entailment patterns when the adjective is modified by half or partially. As shown in (1) and (2), taken from Kennedy and McNally (2005, their examples (39) and (40)), predicating half plus a maximum standard adjective of X doesn't entail that the adjective holds of X, as in (1), whereas predicating half plus a minimum standard adjective of X does entail "adjective of X," as in (2).

(1) a. The plant is half dead. The plant is not dead.

b. The glass is partially full. The glass is not full.

(2) a. The door is half open. The door is open.

b. The table is partially wet. The table is wet.

In their analysis of absolute adjectives with antonyms, Rotstein and Winter (2004) proposed that the standard value of a maximum standard adjective, in their terminology a 'total adjective' (the first member of the pair in clean-dirty, safe-dangerous, healthy-sick), is the lower bound of the corresponding minimum standard ('partial') adjective. In other words, the standard value of clean is the lack of dirt, the standard value of safe is the lack of danger, and so forth. Minimum standard adjectives, on the other hand, may express any point on the scale. (Although

the terminology differs, as far as we can tell, Kennedy & McNally's and Rotstein & Winter's analyses are entirely compatible with each other.)

Adjective phrases with modifiers like slightly appear to be sensitive to whether the adjective has a maximum or a minimum standard. They are less appropriate with maximum standard than with minimum standard adjectives (Rotstein & Winter, 2004). We check this claim in a speeded acceptability judgment experiment (Experiment 1) below. Modifiers like completely are also sensitive to whether the adjective has a maximum or minimum standard. As Kennedy and McNally (2002) show, completely has two meanings: one refers to the end of the scale and the other is synonymous with very (as in The lecture was completely boring). Rotstein and Winter claim that completely gets the former meaning with maximum standard adjectives (completely clean) whereas it gets the latter meaning when combined with minimum standard adjectives if it is acceptable at all (completely dirty). We test this claim in Experiment 2.

All natural languages contain gradable predicates and the means to make comparisons though the syntax of comparatives seems to vary rather widely across languages (see Kennedy, 2006, for an overview). An abstract representation of measurement, or scale, is either explicitly, as in the analysis assumed here, or implicitly (Seuren, 1978) part of the analysis of gradable predicates. For absolute adjectives, we assume, following Kennedy, that a scale is an ordered set of degrees. Our own studies will not focus on potential differences in the measurement functions or ordering relations imposed on scales by particular predicates (e.g, increasing properties like tall versus decreasing properties like short) or differences implied by a particular conceptualization of the property being measured (e.g., is cold conceptualized as a positive or increasing property or as the absence of heat?), or indeed the different formal properties of scales

(e.g., is the scale such that different intervals can be compared in size?). Instead we will investigate the role of the standard value of absolute adjectives and by contrasting the behavior of adjectives with different standard values investigate whether scales are at work in simple comprehension of language (on the assumption that activation of the standard value of a scalar adjective implies the existence/activation of the scale on which the standard is placed). Showing the activation of a particular ordering relation on the scale is another matter - one we find fascinating but not the within the purview of the present paper.

Assuming that there is a general difference between maximum standard and minimum standard adjectives, the issue we wish to address here is whether processing the scales on which the standard is placed is an obligatory part of understanding a sentence that contains a scalar adjective. In the area of syntactic processing, few investigators today doubt that a completely connected syntactic representation of a sentence is built nor that this syntactic analysis is assigned incrementally as the words of a sentence are encountered. Many investigations of syntactic processing have been devoted to the issue of whether some particular type of information influences the incorporation of a new word into the syntactic representation early, when the word is first connected to the syntactic representation for preceding parts of the sentence, or whether it only influences analysis later, e.g., in some stage of reanalysis. In these cases, "later" might still mean within a few hundred milliseconds of encountering a word. In the case of semantic processing, it is not yet known what is part of the obligatory processing of the sentence. How deeply a sentence is semantically processed, which inferences are drawn in which contexts, what interpretive decisions are part of normal comprehension is largely an open matter. Some types of inferences do not seem to be drawn unless they are required (Breheny,

Katsos, & Williams, 2006; McKoon & Ratcliff, 1991; Sanford & Sturt, 2002); other interesting semantic inferences seem to be drawn quickly and influence the continued processing of sentences (Frazier, Pacht, & Rayner, 1999; Piñango, Zurif, & Jackendoff, 1999). So in the case of semantic processing, in addition to the temporal issue concerning precisely when some particular type of semantic information comes into play, there is also the issue of whether using some particular type of semantic information is part of the normal comprehension process in general, or alternatively only in specialized contexts, or only with deliberative judgments, or only when the sentence is understood with a particular task or goal in mind. If processing the standard value/scale of a scalar adjective is part of normal comprehension of the adjective largely independent of whether a metalinguistic judgment is required or of the need to verify a statement, this would begin to set a limit on what information is part of ordinary sentence comprehension, e.g., during the reading of a sentence independent of any post-sentence task or post-sentence deliberations. To address this question, we need to find out whether the effect of scales can be observed during the processing of a sentence before any post-sentence inferencing can take place (see Experiment 3). Another approach to the question is to see whether the standard value/scale of a scalar adjective influences the preferred interpretation of other constituents of its sentence. This too might suggest that the standard value/scale was active during comprehension of the sentence (see Experiment 4).

There has been some investigation of processing scalar adjectives (though relative or 'vague' adjectives are usually given as examples). Clark (1969) studied 'positive' adjectives and found that the positive member tall or long was less complex and more accessible than the 'negative' member short. Gross, Fischer and Miller (1989) studied the lexical organization of

antonyms, claiming that that all predicate adjectives are mentally organized as antonyms either directly or as synonyms to an adjective with a direct antonym. In visual world studies, Sedivy, Chambers, Tanenhaus, and Carlson (1999) have argued that scalar adjectives are used when there is a contrast set. Listeners predict that a contrasting object is the referent when they hear Touch the tall.….

Rips and Turnbull (1988) studied relative (big, small, tall, short) versus absolute (red, green, square, triangular, wooden, American) adjectives in various verification tasks. For example, in their first experiment, an absolute (six-legged) and a relative (small) adjective could be compared in a sentence with a predicate adjective (3) or with a predicate noun (4).

(3) a. An insect is small.

b. An insect is six-legged.

(4) a. An insect is a small animal.

b. An insect is a six legged animal.

Sentences were visually presented for verification. The idea was that relative adjectives may be more context dependent and thus may exhibit longer response times and less accurate responses than absolute adjectives in the predicate adjective condition, where the associated reference class isn't explicitly given, than in the predicate noun condition where it is given. Absolute adjectives should not show this effect. In other words, finding a standard for the relative adjectives may involve extra computation when the reference class isn't mentioned. The results showed that relative adjectives basically behaved as predicted: they were responded to more quickly when the predicate noun was provided than when it was not whereas the absolute adjectives were basically unaffected by whether the sentence contained a predicate noun. However, there was also a

consistency effect: adjectives were verified faster when they satisfied the property of the adjective given an anthropomorphic standard (e.g., a tree would be tall if compared to the height of a human) than if they didn't (e.g., a flower would not be tall if compared to the height of a human). This result bolsters the claim that the scale structure is different for absolute vs. relative adjectives (see also Syrett, Bradley, Kennedy & Lidz, 2005, for evidence from acquisition).

In the present paper, we will be concerned primarily with absolute adjectives, comparing those with a maximum standard (maximum standard adjectives) and those with a minimum standard (minimum standard adjectives).

Experiment 1

The first experiment explores the processing of minimum standard and maximum standard adjectives in an on-line speeded acceptability judgment study. We used (primarily) adjectives taken from Rotstein and Winter (2004) and tested for a difference between the two adjective types by looking at how acceptable they are when modified by slightly or a little. Following Rotstein and Winter we expected that acceptability would be degraded when a maximum standard adjective was modified by slightly (or a little), but this would not be true of minimum standard adjectives.

Before turning to the description of the experiment, two points should be emphasized concerning the goals of the experiment. First, if we want to attribute the behavior of some linguistic item to a particular source, it is important that it behaves as part of a class of items. If we are studying just a single item, it is very difficult to be sure which of its indefinite number of properties is responsible for its behavior. One important goal of Experiment 1 was to determine if indeed maximum standard and minimum standard adjectives differ from each other as a class,

i.e., permitting generalization across items.

A second and equally important goal of Experiment 1 was to make sure that the linguistic intuitions underlying the development of semantic theories of adjectives cannot be attributed to the minimal pair methodology that is standard in linguistics. In at least some domains, it is clear that intuitions obtained in the context of minimal pairs influence in important respects what the data are. For example, in her studies of Japanese intonation, Hirotani (2004) found that matrix questions and embedded questions were produced with distinct intonation patterns only when they were presented together. When embedded questions were pronounced separately from matrix questions, the distribution of intonational patterns was different. Embedded questions were spoken with both the pattern previously attributed to matrix questions and the pattern previously attributed to embedded questions. In other words, what are truly overlapping distributions (indeed, a relation of proper inclusion) appeared as non-overlapping distributions when speakers were presented with the two sentence types in the same block (even without explicit comparison). Similarly, when a sentence like I'd like to know who hid it where when was directly compared for acceptability with a similar sentence that violates a presumed "superiority condition" (see Clifton, Fanselow, & Frazier, 2006, for discussion and data), like I'd like to know where who hid it when, the latter "superiority violation" sentence was rated as acceptable as the former sentence. Presumably, the existence of the phonologically-clumsy sequence where when lowered the acceptability of the sentence that contained it. In a direct comparison with a superiority violation sentence, this clumsiness favored the superiority violation sentence. However, when the superiority violation sentence was presented without the comparison sentence, it was rated worse than the sentence without a superiority violation. In

other words, the existence of a competitor with a flaw influenced the judgment of the superiority violation but only when the two sentences were presented together. This result, like Hirotani's (2004) results, suggests that the minimal pair method may alter our intuitions about language.

In Experiment 1 the sentences were randomized together with filler sentences, and participants never saw both the maximum standard adjective and the minimum standard adjective form of a given sentence. Thus if the penalty for maximum standard adjectives together with a modifier like slightly is observed, it could not be attributed to an unconscious influence of judging minimal pairs.

Experiment 1 collected speeded judgments of the acceptability of sentences like those in (5). Based on Rotstein and Winter's (2004) discussion of maximum standard ("total") and minimum standard ("partial") adjectives, (5b) should be rejected more often than the other forms because in (5b) a maximum standard adjective receives a default value as the lowest value of the corresponding minimum standard adjective and therefore should not be acceptable with a modifier like slightly or almost, at least not on its default interpretation. Assuming that the predicted default interpretation is assigned first, before any 'coerced' or non-default interpretation, then longer reaction times might also be expected for the "acceptable" responses in (5b) due to the extra operations needed to establish a non-default interpretation. In Kennedy and McNally's (2005) terms, the maximum standard of the default interpretation of the maximum standard adjectives will presumably conflict with the presence of a minimizer like slightly or a little but the default interpretation of a minimum standard adjective will not.

Methods

Materials. Sixteen sentences like (5) were constructed with four versions of each.

(5) a. maximum std. I would say that this table is clean.

b. slightly/maximum std. I would say that this table is slightly clean.

c. minimum std. I would say that this table is dirty.

d. slightly/minimum std. I would say that this table is slightly dirty.

Two (5a, b) contained a maximum standard adjective; two (5c, d) contained the corresponding minimum standard adjective. Two (5b, d) contained a modifier (slightly in half the items, a little in the other half); two contained no modifier (5a, c). All sentences appear in Appendix A. The critical clause (this table is clean) always appeared as a complement clause simply in order to lengthen the sentences and to keep them from standing out from the others in the experiment. These 16 sentences were combined with a total of 120 other sentences of various forms (including verb phrase ellipsis and other ellipsis sentences of varying acceptability and 36 clearly unacceptable filler sentences) following 8 varied practice sentences. Four counterbalanced forms of the list were constructed, each with 4 experimental items in each version.

Participants and procedures. Sixty University of Massachusetts undergraduates were tested in individual half-hour sessions. Each participant was first instructed that the task was to "decide, just as quickly as possible, whether [each sentence they saw] is good, grammatical, sensible, and meaningful." The participant was to pull a trigger with the right hand if a sentence was acceptable in the sense just described, and a trigger with the left hand if it was not. Following instructions, they saw the eight practice items (half clearly unacceptable for various reasons, e.g., missing arguments) and then saw the 136 items in an individually-randomized order. Each item appeared in two separate parts on a computer monitor. First, a display appeared on the monitor with underscores indicating where letters would be. The participant pulled a trigger with the right

hand to replace the first line of underscores with the matrix "lead-in" to the sentence (I would say that). After reading this, the participant pulled the trigger again, which replaced this lead-in with underscores and brought up the critical embedded clause on the second line. The participant indicated the acceptability of the sentence containing this clause quickly by pulling the right-hand or left-hand trigger, as instructed. The computer that controlled the experiment recorded choice and reading time for each presentation region of the sentence. Trials with times over 8000 ms were discarded (1.5% of all trials).

Results

Table 1 presents the mean proportion of "acceptable" responses and the mean reaction times to the second presentation region of the sentences (averaged over both "acceptable" and "unacceptable" responses, since there were too many missing data of each type in Condition 5b to analyze the two types of responses separately; further, the reading time data for one participant had to be eliminated because of an excessive number of long reaction times). Analyses of variance of the proportion of "acceptable" responses indicated significant main effects of each factor (type of adjective, $F1(1,59) = 15.20$; $F2(1,15) = 13.43$) and presence of slightly/a little, $F1(1,59) = 32.14$; $F2(1,15) = 22.78$) and the interaction between them ($F1(1,59) = 17.94$; $F2(1,15) = 12.24$; all $p < .01$). All effects reflected the infrequent judgments of "acceptable" for modified Maximum standard adjective items like (5b).

The mean reaction times appeared to show the same pattern, but the long decision times for modified maximum standard adjectives resulted only in two significant main effects but not a significant interaction (maximum standard adjectives longer than minimum standards, $F1(1, 58) = 15.77$; $F2(1,15) = 23.03$), modified longer than unmodified, $F1(1,58) = 21.54$; $F2(1,15) = 9.15$;

p always < .01, but interaction F1(1,58) = 2.21, p < .15; F2(1,15) = 1.30, p = .27). While the reaction time data were apparently too variable to place much faith in, they do not contradict the evidence from proportion "acceptable" showing that maximum standard adjectives modified with slightly or a little pose a problem.

Discussion

The percentage acceptable responses clearly supported the Rotstein and Winter (2004) prediction: adding a modifier like slightly decreased the acceptability of maximum standard adjectives but not minimum standard adjectives. The response times showed a similar if not convincingly significant disadvantage for maximum standard adjectives modified by slightly or a little.

The results suggest that the standard for maximum standard and minimum standard adjectives do differ as described above, following Rotstein and Winter's analysis. A maximum standard adjective, such as dry or clean, receives as a default value the lowest value of the corresponding minimum standard adjective (wet, dirty). Modification with slightly or a little requires an interpretation as a nonmaximal degree, which is inconsistent with the default value, leading to questionable acceptability. A minimum standard adjective, on the other hand, by default indicates some non-zero degree, fully consistent with modification by slightly and therefore fully acceptable when modified. On the Kennedy and McNally analysis, the maximum standard itself may conflict with a modifier that is a minimizer, such as slightly or a little.

The results are encouraging for current semantic accounts of absolute scalar adjectives. The theories were supported even with speeded judgments and under circumstances where a participant who saw a particular adjective did not also see its antonym. This reinforces the

conclusion that maximum standard and minimum standard adjectives are behaving as classes.

Whether these classes are fully homogeneous remains unclear. One distinction among adjectives that has been drawn in the literature is based on whether the adjective pair is "complementary" or not. Cruse (1980) introduced this notion, using it to apply to adjectives if, whenever the property denoted by adjective 1 or by its antonym adjective 2 could apply, one of the two properties must apply, e.g., an object is either clean or it is dirty. According to Cruse, all of the adjectives tested in Experiment 1 would be classified as complementary. However, Rotstein and Winter (2004, Table 1) distinguish among the adjectives tested in Experiment 1, classifying only a subset of them, the ones with negative prefixes, as complementary, and noting that the others could be complementary or not depending on context. They also note that with complementary adjective pairs, modifying the maximum standard adjective with almost is acceptable (almost complete) but modifying the minimum standard adjective is not (#almost incomplete). However, they also claim that modifying their noncomplementary minimum standard adjectives with almost can be felicitous depending on context (e.g., almost hungry, almost dirty, almost sick in contexts with minor thoughts of food, minor specks of dirt, or minor symptoms).

Using the classification in Rotstein and Winter's Table 1 as a basis for the classification, the adjectives from Experiment 1 were divided into 'complementary' and 'non-complementary' pairs to determine if the slightly-penalty was carried entirely by the complementary adjectives. This proved not to be the case. While the penalty for modifying a minimum standard adjective with slightly (measured as the interaction in proportion "acceptable" judgments between maximum vs. minimum standard adjective and presence vs. absence of slightly) was numerically

greater for complementary than noncomplementary adjectives (0.37 vs. 0.23), the difference between these two values was not significant ($t(12) = 1.18$, $p > .25$).

Experiment 2

The second experiment was designed to provide more evidence that the standard of maximum standard and minimum standard adjectives differ. Consider (6) and (7) (taken from Rotstein & Winter, 2004; their example (40)). Rotstein and Winter propose that the (a) forms are less acceptable than the (b) forms. The (a) forms have a maximum standard adjective modified by completely, followed by a statement that expresses a degree comparison in terms of a minimum standard adjective. In the (b) forms, the appearance of the adjectives is reversed. The adjective completely is an overt marker of the complete scale. This should reinforce the default maximal interpretation of a maximum standard adjective, which makes it odd (in the (a) forms) that the speaker should go on and assert something incompatible with the maximum standard adjective having a value at the very end of the scale. However, the modifier completely when applied to a minimum standard adjective can have an interpretation more like very, in which case a comparative statement (expressed in terms of a comparative form of a maximum standard adjective) would be acceptable.

(6) a. #The two towels are completely dry, but the red one is a little bit wetter than the blue one.

b. The two towels are completely wet, but the red one is a little bit drier than the blue one.

(7) a. # The kitchen and living room are completely clean, but the kitchen is a little bit dirtier.

b. The kitchen and living room are completely dirty, but the kitchen is a little bit cleaner.

Rotstein and Winter comment that "for an antonym pair A and B, completely A means no amount of B if a zero amount of B is meaningful." (2004, page 281). When A is a maximum standard adjective like dry, a zero amount of its minimum standard antonym B (wet) is clear. It is simply the default meaning of the maximum standard adjective. It would be inappropriate to differentiate two entities, each of which has a zero value, as in (6a). However, when A is a minimum standard adjective (wet), a zero amount of its maximum standard counterpart (dry) is not clear. No lower bound is part of the standard value of the maximum standard adjective. Therefore, completely A does not mean zero B when A is a minimum standard adjective, leaving it possible to compare two "completely A" entities in terms of the degree to which they exhibit B.

In Experiment 2, examples like (6) and (7) were tested to check whether the relatively subtle intuition noted by Rotstein and Winter is general and appears in a speeded acceptability judgment task like that used in Experiment 1.

Methods

Materials. Sixteen sentence pairs like those in (8) were constructed based on Rotstein and Winter's example (their 40 = our 6). Each sentence had two clauses connected by but. The first clause asserted that two entities of some type (towels) are completely or absolutely [adjective]. In the a-form the adjective was the maximum standard adjective and the b adjective was the minimum standard adjective. In the continuation clause, one of the entities was described as being higher on the named scale than the other entity. For the a-form, the continuation included the minimum standard adjective in its comparative form; for the b-form, the continuation

contained the comparative form of the maximum standard adjective. All materials appear in Appendix 2.

(8) a. These juices are absolutely pure, but the mango juice is more impure than the papaya.

b. These juices are absolutely impure, but the mango juice is purer than the papaya.

These 16 experimental items were combined with 132 other items of a variety of constructions (including ellipses, superiority violation sentences, conjoined clauses, and quantified reciprocals; the 36 clearly unacceptable filler sentences used in Experiment 1 also were used in Experiment 2). Two counterbalanced lists were constructed, with half of the experimental items appearing in each version (8a or 8b) in each list.

Participants and procedures. Forty-eight University of Massachusetts undergraduates were tested in individual half-hour sessions. The procedures were the same as those described in Experiment 1, except that a separate 6-item practice list followed by a short break preceded the experimental list. The first presentation segment of each sentence was its first clause, up to the comma, and the second presentation segment, which appeared on the second line of the display, was the clause with the comparative statement in it. As in Experiment 1, participants were instructed not to reject a sentence until reading its second segment.

Results

Table 2 presents the mean percentages of "acceptable" responses and the mean reading times for both the first and the second presentation segments (pooled over both "acceptable" and "unacceptable" judgments, as in Experiment 1). As predicted, participants accepted significantly fewer items with a modified maximum standard adjective followed by a comparative minimum standard adjective (like 8a) than they did items with the minimum standard-maximum standard

order (like 8b) ($F1(1,47) = 26.78$; $F2(1,15) = 12.93$; $p < .01$). While reading times for initial presentation segments with a modified maximum standard adjective were numerically faster than for initial segments with a modified minimum standard adjective, the difference was nonsignificant ($F1(1,47) = 2.08$, $p > .15$; $F2(1,15) = 1.05$, $p > .30$). Similarly, the apparent difference in reading times for the second presentation segment was nonsignificant (all $F < 1.0$). We doubt that these nonsignificant reading time differences should be given any credence, since they are opposite to the significant proportion acceptability judgments we obtained.

Discussion

As expected, the maximum standard adjectives followed by what might be called a disclaimer continuation were judged less acceptable than their minimum standard adjective counterparts. This supports the basic distinction among the adjectives, including in rather subtle examples. If a maximum standard adjective denotes the absence of the property denoted by its antonym, in Rotstein and Winter's terms, or it has a maximum standard, in Kennedy and McNally's terms, then it should be odd to assert that an entity is, say, clean and then disclaim that assertion by claiming it has more dirt than some other entity. Thus the results support a semantic analysis of these adjectives that draws a distinction among types of absolute adjectives and they fit well with a degree semantics in which maximum standard adjectives specify a (default) standard that is the maximum on a closed scale.

Experiment 3

An important question to ask about the topic of the present research is whether properties of scales, in particular the minimum vs. maximum standard aspect of the semantic denotation of a word, are obligatorily processed as part of comprehending the meaning of a phrase or sentence

containing a scalar adjective. We dub the hypothesis that they are the "Obligatory Scale" hypothesis. It is possible that the distinction between maximum standard and minimum standard adjectives becomes apparent to a reader only upon reflection, as when an explicit judgment of acceptability is required (as it is in Experiment 1) or when a contrast with the opposite polarity term is required (as in Experiment 2). While it is generally thought that syntactic processing is obligatory (but cf. Ferreira, 2003, for some reservations), and there is evidence that some semantic processing is obligatory and even quick (e.g., Rayner et al, 2004), it may be that some semantic inferences are delayed until they are specifically required (McKoon & Ratcliff, 1992). This appears to be the case for instrumental inferences (e.g., McKoon & Ratcliff, 1981) and predictive inferences (McKoon & Ratcliff, 1986). It may also be the case for one kind of pragmatic scalar implicature, e.g., some => not all. These implicatures do not seem to be obligatorily drawn when one has understood a word (some) that can be placed on an entailment scale: all > many > some (Breheny et al., 2006, Chierchia et al., under revision). Drawing them requires cognitive resources that readers may not normally commit.

Thus, finding that the distinction between maximum standard and minimum standard adjectives is regularly honored in the reading of normal text would be of interest. If the Obligatory Scale hypothesis is correct, then there should be a disruption in the eye movement record during normal reading, indicating longer processing times, for sentences in which the default meanings of scalar adjectives are not honored (see Rayner, 1998, and Clifton, Staub, & Rayner, in press, for surveys of eyetracking studies showing effects of comprehension difficulty). In particular, sentences in which slightly or a little modifies a maximum standard adjective should produce eye movement disruptions. Although the Obligatory Scale hypothesis

does not make precise predictions about exactly when the disruption should become apparent, the hypothesis clearly predicts that it should be during the processing of the sentence, not as part of some post-sentence assessment computation.

To investigate this question, we conducted an eye movement recording experiment testing sentences adapted from the materials used in Experiment 1, but placing the critical phrase so that it appeared well before the end of the sentence so that the region immediately following the critical adjective wouldn't coincide with the end of the sentence. Rather than contrasting modifier and no modifier conditions, as in Experiment 1, the modifier slightly (or a little) was compared to the modifier completely, which should be unproblematic with either a maximum standard or a minimum standard adjective. The reason for comparing the two modifier conditions rather than a modifier and unmodified condition was in order to keep the length of constituents more comparable across conditions.

Methods

Materials Twenty experimental sentences were constructing using the modifiers and adjectives from Experiment 1, with four versions of each, as in (9). The a-form contained either slightly or a little and a maximum standard adjective; the b-form substituted the corresponding minimum standard adjective but was otherwise identical to the a-form. The c-form in all cases contained the adverb completely and the maximum standard adjective; the d-form was identical to the c-form apart from substituting the corresponding minimum standard adjective for the maximum standard adjective.

(9) a. This table is slightly clean right now, and the new bus boy is clearly responsible. He seems very young, this new guy. (Slightly maximum standard)

b. This table is slightly dirty right now, and the new bus boy is clearly responsible. He seems very young, this new guy. (Slightly minimum standard)

c. This table is completely clean right now, and the new bus boy is clearly responsible. He seems very young, this new guy. (Completely maximum standard)

d. This table is completely dirty right now, and the new bus boy is clearly responsible. He seems very young, this new guy. (Completely minimum standard)

All of the sentences were followed by a continuation sentence, as in (9). Six of the sentences were followed by a relevant yes/no question, e.g., Is this table clean now? NO....YES

The twenty experimental items were combined with 96 other items of a wide variety of constructions, preceded by eight practice items. Four separate counterbalancing lists were constructed, in each of which five experimental items appeared in each of the four versions illustrated in (9).

Participants and procedures. Thirty-six University of Massachusetts undergraduates participated in individual 50-minute sessions, for pay or course credit. The sentence stimuli were presented as two lines (as indicated in (9), with the line break occurring before the end of the first sentence, before responsible in this example) on a monitor controlled by a Pentium PC. An A to D converter interfaced the computer with a Fourward Technologies Generation V Dual Purkinje eyetracker. The eyetracker monitored movements of the right eye, although viewing was binocular. Letters were formed from a 7 X 8 array of pixels, using the Borland C default font. Participants sat 61cm away from a computer screen and silently read single line sentences while their head position was stabilized by a bite bar. At this viewing distance, 3.8 letters equaled one degree of visual angle. At the beginning of the experiment, the eye-tracking system

was calibrated for the participant. At the start of each trial, a check calibration screen appeared, and participants who showed a discrepancy between where their eye fixated and the location of the calibration squares were re-calibrated before the next trial. A trial consisted of the following events: The check calibration screen appeared and the experimenter determined that the eye-tracker was correctly calibrated. The participant was instructed to look at the calibration square on the far left of the screen, which triggered the presentation of the sentence. The participant read the sentence silently and at his/her own pace, then clicked a response key to make the sentence disappear. Following 1/4 of the trials, a comprehension question appeared on the screen. The participant responded by pressing the response key that corresponded with the position of the correct answer. Then the check calibration screen appeared before the next trial. On the first eight trials of an experimental session, practice sentences were presented. Following that, the order of presentation of sentences was individually randomized for each participant.

Results

As is the normal practice in eye movement research, the data were analyzed in a wide variety of ways. For brevity, only the most informative analyses will be presented here. Other conventional ways of analyzing eyetracking data either yielded null results or were redundant with the results reported here.

In analyzing the data, the materials were divided into regions as indicated by the / marks, subscripted for region number, in (10) (note, the line break occurred after Region 4):

(10) This table is/$_1$ slightly clean/$_2$ right now,/$_3$ and the new bus boy is clearly/$_4$ responsible./$_5$
He seems very young, this new guy./$_6$

The means of the most informative reading time data for each region of the sentences, through the end of the first line (region 4) appear in Table 3. These data were subjected to analyses of variance, treating both participants (F1) and items (F2) as random variables, and treating counterbalancing participant and item groups as factors (cf. Pollatsek & Well, 1995).

An immediate response to the clash between slightly or a little and a maximum standard adjective would appear as longer reading times in region 2. This did not happen. The "first pass" reading times for region 2 (the sum of all fixation durations in the region from first entering it until first leaving it) did not differ among conditions. Most saliently, the predicted interaction between adverb (slightly vs. completely) and adjective (maximum standard vs. minimum standard) in region 2 was not significant, $F < 1$. An analysis of regions 2-4 (treating regions as a factor) yielded similarly nonsignificant results.

An analysis of the second-pass reading times (the summed fixation times in each region that occurred after the region was exited or gone past) gave a marginally significant indication of the predicted difficulty for maximum standard adjectives modified by slightly. In an analysis of second-pass times for regions 1-3, treating regions as a factor, the interaction between type of adverb and type of adjective approached significance ($F1(1,32) = 3.734$, $p = 0.062$; $F2(1,16) = 3.389$ $p = 0.084$), and second pass times for items with slightly were significantly longer than for items with completely ($F1(1,32) = 6.37$, $p < .02$; $F2(1,16) = 8.91$, $p < .01$). An analysis of the total reading times (the sum of all fixation durations in a region) for regions 1-3 yielded a fully significant interaction of type of adverb and type of adjective ($F1(1,32) = 10.290$, $p = 0.003$; $F2(1,16) = 7.155$, $p = 0.017$) in addition to significant or nearly-significant main effects of both

type of adverb and type of adjective (reflecting the long total times for maximum standard adjectives modified by slightly).

The second pass and total times do reflect the predicted difficulty of sentences like (9a). However, the first pass times did not. It is possible that the observed difficulty resulted completely from trials on which participants reached the end of the entire item and then re-read the difficult parts. This would not be strong evidence for the Obligatory Scale Hypothesis that motivated Experiment 3, but could result from "ruminative" processing that takes place after the sentence was initially comprehended. However, evidence for a processing effect during the initial reading of a clause does appear when the entire first line was treated as one region, and the first pass time for this region was calculated. These first pass times include all the time spent reading the first line before the eyes had gone on to line 2, but do not contain any 'post-rumination' time that might have contributed to the total time effects reported earlier. The mean first line first pass reading time was 2679 ms for items like (9a), maximum standard adjectives modified by slightly, while it was 2499, 2493, and 2499 for items like (9b), (9c), and (9d), respectively. The interaction between adverb and type of adjective was fully significant ($F1(1,32) = 6.477$, $p = 0.016$; $F2(1,16) = 7.013$, $p = 0.018$).

Discussion

There was clear evidence of disrupted reading in the condition that was predicted to be difficult, the condition with slightly or a little plus a maximum standard adjective. The disruption did not appear in initial reading of the critical adverb + adjective region, but did appear in the second pass reading times (marginally) and total reading times for the first three regions of the items. Further, and most informatively, it appeared in the first pass reading times for the first line

of the two-line items we used. These times include time spent re-reading the critical adverb + adjective region plus adjoining regions, but do not include any reading time that occurred after the end of the first line was reached.

We conclude that modifying an adjective with an adverb that clashes with the default meaning of the adverb disrupts reading. This disruption does not appear as immediately in the eyetracking record as the effects of lexical variables (e.g. word frequency) or some syntactic variables do (see Clifton, Staub, & Rayner, in press, for some discussion), but it does appear during the normal reading of a sentence, prior to reaching the end of the sentence or the text containing it. Further research is needed to probe into just how quickly the clash between adverb and default meaning of the adjective is detected. It is possible that the language processing system detects it quickly but the eyes do not react to it until later in the clause, but it is equally possible that the clash is simply not detected until a substantial amount of sentence processing has been completed. Nonetheless, we propose that Experiment 3 supports what we have called the Obligatory Scale hypothesis as opposed to possible hypotheses that the 'standard' values of absolute scalar adjectives are honored only when the context forces it, as appears to be the case with several types of semantic information discussed earlier.

Experiment 4

Experiments 1-3 have demonstrated that combinations of adverb+adjective are difficult to process or are judged relatively unacceptable when the adverb is inappropriate for the default meaning of the adjective, where the default meaning is specified by whether the adjective is a maximum standard or minimum standard absolute scalar adjective. Experiment 4 is designed to see if the difference between maximum and minimum standard adjectives has any effect on the

meaning typically assigned to a sentence when the adjective is modified by a quantifying adverb. Such a finding would demonstrate that the maximum/minimum standard difference contributes to the compositional understanding of a sentence's meaning, not just to the sentence's difficulty.

Adverbs may quantify over occasions, as in (11a), or over individuals, as in (11b).

(11) Students usually walk to school.

a. On occasions when they go to school, usually they walk.

b. Most students walk to school.

Now consider what happens in examples like (12), in which mostly modifies a maximum standard or a minimum standard adjective. Each can receive either a 'degree' interpretation, in which mostly specifies a value on the adjective's scale, or a 'quantity' interpretation, where mostly quantifies over individuals.

(12) a. The dishes were mostly clean.

b. The dishes were mostly dirty.

'Degree' paraphrase: The dishes were clean/dirty to a large degree.

'Quantity' paraphrase: Most of the dishes were clean/dirty.

Intuitively, (12a), containing a maximum standard adjective, seems more likely to receive a degree interpretation than (12b), with a minimum standard adjective. (12b) seems more likely to receive a quantity interpretation than the maximum standard adjective does. This intuition is explained by the assumption that maximum standard adjectives like clean have as their default interpretation a maximum standard, let's call that 100%. If we understand clean as 100% clean, then we can understand mostly clean as most of the way to 100% clean, e.g. 80% clean. In sentences like (12a), quantifying over the scale associated with the maximum standard adjective

should thus be unproblematic. By contrast, minimum standard adjectives like dirty may not introduce a clear standard that may be modified by mostly. If we think of a minimum standard as some non-zero amount, then the standard introduced by the default interpretation of the minimum standard adjective is not what we need to interpret mostly, i.e., mostly dirty doesn't mean most of the way to some non-zero amount of dirt. If readers are unsure what degree should be modified by mostly they may be more likely to look for an alternative (non-degree) interpretation such as one where mostly quantifies over the domain of individuals. If this line of reasoning is correct, then readers may choose fewer degree paraphrases for minimum standard adjectives than for maximum standard adjectives because only the standard of the default interpretation of the adjective is required in computing the degree interpretation for maximum standard adjectives but not for minimum standard adjectives.

Using a computerized paraphrase selection task, Experiment 4 tested the hypothesis that maximum standard and minimum standard adjectives differ in how accessible a degree interpretation is.

Methods

Materials. Twelve pairs of sentences like those in (12) were constructed, one (the a form) with a maximum standard adjective and one (b) with a minimum standard adjective. All adjectives were modified with mostly. All experimental materials appear in Appendix 4. A two-choice interpretation question was made up for each sentence, as illustrated following (12). In the questionnaire, the first of the two options expressed the quantity interpretation of the sentence, and the second, the degree interpretation.

Participants and procedures. Forty-eight University of Massachusetts undergraduates participated in individual half-hour sessions. Two counterbalanced lists were created, in each of which six of the 12 sentences appeared with each type of adjective, maximum standard vs. minimum standard. These 12 sentences were combined with 29 other sentences of a variety of forms that required a choice between two interpretations, and 73 other sentences of various forms that required an acceptability rating on a 5-point scale. A computer randomized the resulting 114 sentences for each subject and presented the sentences and the response request on a video monitor. Sentences for which a choice of interpretation was required were preceded by the presentation of an instruction to comprehend the following sentence and prepare to select one of two interpretations that would appear on the next screen. Sentences for which an acceptability rating was required were preceded by an instruction to read and rate the following sentence. The computer recorded the choices the participant made, but did not record response times. standard adjective modified by mostly received 57.2% degree interpretations and only 42.8% quantity interpretations. Sentences containing a modified minimum standard adjective received 39.9% degree interpretations and 60.1% quantity interpretations. The difference in frequency of interpretations was significant ($t1(47) = 4.90$, $p < .001$; $t2(11)= 3.24$, $p < .01$).[2]

Discussion

The results supported the prediction that more degree interpretations would be assigned

[2] Louise McNally (p.c.) noted that some of our items might prefer an interpretation in which quantification takes place over time or events, not over 'quantity' or degree, and suggested that this preferred interpretation might affect the choice between the two other interpretations offered. We identified three such items (truthful, satiated, healthy) and examined them separately. These three items as well as the remaining nine items showed more frequent degree vs. quantity interpretations for maximum than for minimum standard adjectives, 0.26 for the

when mostly modified a maximum standard adjective, with a clear value for what would count as having the relevant property completely or 100%, than when mostly modified a minimum standard adjective, where the minimum standard could not be modified by mostly. The results show very strongly that the scale and standard introduced by the semantics of a scalar adjective sentence is computed as part of the normal processing of the sentence. Judgments about the acceptability of the sentence are not required. The scale may have an indirect effect by influencing the most tempting domain of quantification in sentences like those tested here.

In other contexts, too, distinct interpretations seem to arise for maximum vs minimum standard adjectives (see discussion in Yoon, 1996). There seems to be an interaction of maximum standard vs. minimum standard adjective with the maximality presupposition (assuming that the carries both an existence and a maximality presupposition). Consider (13), uttered when your sloppy teen-age son has just come in and dropped a pile of clothes on the floor.

(13) a. Put your dirty clothes where they belong!

b. Put your clean clothes where they belong!

The former sentence (13a) has two construals. It has a strong interpretation, where it refers to just the clothes that are actually dirty, and a weak interpretation, where it refers to the entire pile of clothes. The latter sentence (13b) has only the strong interpretation. Only the clean clothes are to be put in away.

The quantificational facts above concerning mostly, and the observation about definite plural phrases in (13) both suggest that the contrast between adjectives with a maximum standard

---

three temporal-quantification adjectives vs. 0.15 for the rest.

(clean) and a minimum standard (dirty) is related to different default meanings for the two types of adjective: a standard that is by default maximal, 100% and without exception for maximum standard adjectives and a standard which is weaker, at no fixed value and tolerating exceptions for minimum standard adjectives.

General Discussion

To summarize, Experiments 1 and 2 show that for at least one set of antonyms involving maximum standard and corresponding minimum standard absolute adjectives, they act as a class. Treating each adjective or adjective pair on its own is not required. Many properties of the members of each class are predictable, e.g., the effects of modification with slightly as claimed by Kennedy and McNally (2005) and by Rotstein and Winter (2004). This is true for acceptability judgments that are obtained without explicit comparison of an adjective and its antonyms and without presenting items as minimal pairs.

Experiments 3 and 4 show that the distinct behavior of these classes of adjectives is observed in normal reading and in paraphrase selection, in the absence of any reasoning task. The eye movement experiment shows that the differences between these classes of adjectives begins to be apparent ion the line of text that includes the adjective, eliminating any theory where the inferences involved with these scalar adjectives is part of a late deliberative reasoning process.

The results of Experiment 4 on mostly further strengthen this conclusion and show that the difference between maximum and minimum standard scalar adjectives contributes compositionally to the meaning of quantified sentences. In a paraphrase selection task where the degree interpretations and (what we have called) ‘quantity’ interpretations can be imagined for

both types of adjectives, we nevertheless see a difference that can be related to the maximum standard of a maximum standard adjective. These results support the Obligatory Scale Hypothesis, suggesting that semantically introduced scales are part of the obligatory computation of the meaning of the sentence. Further evidence that the scale and the standard matters comes from the strong interpretations required for maximum, but not minimum, standard adjectives in definite plural phrases (see (13), above).

Kennedy (to appear) proposed the Interpretive Economy condition in (14).

(14) Interpretive Economy

Maximize the contribution of the conventional meaning of the elements of a constituent to the computation of its meaning.

In Experiment 4, when mostly modified a maximum standard adjective, the majority of responses indicated a degree interpretation. As noted earlier, with maximum standard adjectives there was no need for readers to compute a 'quantity' interpretation where mostly quantifies over the domain of individuals introduced by the subject of the sentence. The preponderance of degree interpretations assigned to the maximum standard adjectives could be taken to support the implicit 'locality' preference in (14). Assuming any sort of roughly bottom-up composition mechanism, (14) will apply to maximize the conventional meaning of words in smaller (local) constituents, such as mostly clean, before it applies to maximize the contribution of the conventional meaning of more distant constituents. We already had an explanation for the difference between maximum and minimum standard adjectives in Experiment 4, but (14) helps to explicitly capture the mechanism that would result in maximum standard adjectives not just receiving more degree interpretations than minimum standard adjectives but a preponderance of

degree interpretations.

Processing the scale structure and the standard seems to be part of the obligatory computation involved in language comprehension, regardless of the particular task at hand, given that it influences the chosen domain of quantification in cases where it need not play a role at all. This can be contrasted with various kinds of inferences (McKoon & Ratcliff, 1992), together with scalar inferences that are introduced by the pragmatics. For example, in non-downward entailing contexts, Breheny et al. (2006) presented evidence that scalar inferences (e.g., for the exclusive interpretation of or) tend to be drawn when preceding context makes them relevant but less often in neutral contexts. Similarly, Chierchia et al. (under revision) presented evidence that fewer scalar inferences involving pragmatic enrichment are drawn in downward entailing contexts than in non-downward entailing contexts. This contrasts with what we find concerning the nature of the scales and standards introduced by the scalar adjectives investigated here. Without any preceding context to trigger scalar inferences, the semantic scales introduced by the adjectives seem to be considered in the interpretation of a sentence containing the adjective. That semantically introduced scales and pragmatically introduced scales should differ is expected in current semantic and pragmatic theories; it is reassuring that these differences also receive psycholinguistic support from a variety of judgment, reading, and interpretation tasks.

References

Breheney, R., Katsos, N., and Williams, J. 2006. Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100 (3), 389-433.

Clark, H. 1969. Linguistic processes in deductive reasoning. *Psychological Review*, 76, 387-404.

Chierchia, G., Frazier, L., and Clifton, C., Jr. (under revision). When basic meanings are (not) enough: Processing scalar implicatures.

Clifton, C. Jr., Fanslow, G., and Frazier, L. 2006. Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry*, 37, 51-68.

Ferreira, F. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164-203.

Frazier, L., Pacht, J. M., & Rayner, K. 1999. Taking on semantic commitments, II: Collective versus distributive readings. *Cognition, 70,* 87-104.

Gross, D., Fischer, U., and Miller, G. A. 1989. The organization of adjective meanings. *Journal of Memory and Language*, 26, 92-106.

Hirotani, M. 2004. Prosody and LF interpretation: Processing Japanese wh-questions. University of Massachusetts doctoral dissertation.

Kennedy, C. 1999. *Projecting the Adjective: The syntax and semantics of gradability and comparison*. New York: Garland.

Kennedy, C. 2006. Semantics of comparatives. In K. Brown (Ed.), Encyclopedia of Language and Linguistics. Second Edition. New York: Elsevier.

Kennedy, C. (to appear). Vagueness and grammar: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy*

Kennedy, C. and McNally, L. 1999. From event structure to scale structure: Degree modification in de-verbal adjectives. In T. Matthews and D. Strolovitch (eds) *SALT IX*, 163-180.

Kennedy, C. and McNally, L. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81, 345-381.

McKoon, G., & Ratcliff, R. 1981. The comprehension processes and memory structures in instrumental inference. *Journal of Verbal Learning and Verbal Behavior, 20*, 671-682.

McKoon, G., & Ratcliff, R. 1986. Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*, 82-91.

McKoon, G., & Ratcliff, R. 1992. Inference during reading. *Psychological Review, 99*, 440-466.

Piñango, M. M., Zurif, E., & Jackendoff, R. 1999. Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research, 28*, 395-414.

Pollatsek, A., & Well, A. D. 1995. On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 785-794.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. 2004. The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory and*

Cognition, 30, 1290-1301.

Rips, L., and Turnbull, W. 1980. How big is big? Relative and absolute properties in memory. *Cognition,* 8, 145-174.

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

Sanford, A.J., and Sturt, P. 2002. Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Science*, 6 (9), 382-386.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., and Carlson, G.N. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109-147.

Seuren, P. A. 1978. The structure and selection of positive and negative gradable adjectives. In D. Farkas and K. Todrys (Eds), *Papers from the parasession on the lexicon* (pages 336-346). Chicago: Chicago Linguistic Society.

Syrett, K., Bradley, E., Kennedy, C., and Lidz, J. 2005. Shifting standards: Children's understanding of gradable adjectives.

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

Acknowledgments. We are extremely grateful to Chris Kennedy for extensive discussion, suggestions and many insightful observations, and to Louise McNally and the participants of the Workshop on Scalar Meaning held at the University of Chicago, May 19-20, 2006 for helpful comments. Author Stolterfoht is now at the University of Tuebingen. This research was supported in part by Grant HD-18708 to the University of Massachusetts.

Appendix 1. Materials for Experiment 1.

1. I would say this table is (slightly) clean|dirty.

2. According to the authorities, this building is (slightly) safe|dangerous.

3. It seems obvious that this student is (slightly) healthy|sick.

4. From time to time, this dog is (slightly) dry|wet.

5. In my opinion, this report is (slightly) complete|incomplete.

6. Janet suggested that this juice is (slightly) pure|impure.

7. The teacher observed that this paragraph is (slightly) clear|unclear.

8. Susie noticed that this stone is (slightly) smooth|rough.

9. According to the brochure, this relic is (a little) whole|cracked.

10. It seems to us that this dog is (a little) satiated|hungry.

11. From our perspective, this road is (a little) straight|curved.

12. At the end of the play, this door is (a little) closed|open.

13. According to the dentist, these dentures are (a little) perfect|imperfect.

14. In the final analysis, this reporter is (a little) truthful|untruthful.

15. From an American perspective, this guide is (a little) certain|uncertain.

16. The biologist knew that this caterpillar is (a little) dead|alive.

Appendix 2. Materials for Experiment 2.

1. The two towels are completely dry, but the red one is a little bit wetter than the blue one.|The two towels are completely wet, but the red one is a little bit drier than the blue one.

2. The rooms are both completely clean, but the one on the left is dirtier than the one on the right.|The rooms are both completely dirty, but the one on the left is cleaner than the one on the

right.

3. Our cars are both completely safe, but my husband's is more dangerous than mine.|Our cars are both completely dangerous, but my husband's is safer than mine.

4. The two visitors are completely healthy, but the one from Japan is sicker than the one from China.|The two visitors are completely sick, but the one from Japan is healthier than the one from China.

5. These reports are absolutely complete, but Ana's is more incomplete than Paula's.|These reports are absolutely incomplete, but Ana's is more complete than Paula's.

6. These juices are absolutely pure, but the mango juice is more impure than the papaya.|These juices are absolutely impure, but the mango juice is purer than the papaya.

7. The term papers are both absolutely clear, but Stacy's is more unclear than Stan's.|The term papers are both absolutely unclear, but Stacy's is clearer than Stan's.

8. Both sculptures are absolutely smooth, but the one there is rougher than the one over here.|Both sculptures are absolutely rough, but the one there is smoother than the one over here.

9. These pots are both completely whole, but this one is more cracked than that one.|These pots are both completely cracked, but this one is more whole than that one.

10. Both guests are completely satiated, but Nick is hungrier than Ryan.|Both guests are completely hungry, but Nick is more satiated than Ryan.

11. The roads are both completely straight, but the highway is more curved than the interstate.|The roads are both completely curved, but the highway is straighter than the interstate.

12. The two doors are completely closed, but the front door is more open than the side door.|The two doors are completely open, but the front door is more closed than the side door.

13. The poems are absolutely perfect, but the new one is more imperfect than the old one.|The poems are absolutely imperfect, but the new one is more perfect than the old one.

14. Both children are completely truthful, but the boy is more untruthful than the girl.|Both children are completely untruthful, but the boy is more truthful than the girl.

15. The advisor was absolutely certain, but the experienced one was more uncertain than the novice.|The advisor was absolutely uncertain, but the experienced one was more certain than the novice.

16. The lizards are completely dead, but the long one is more alive than the short one.|The lizards are completely alive, but the long one is more dead than the short one.

Appendix 3. Materials for Experiment 3. Options separated by $ and | symbols. Sample question illustrated for item 1.

1. This table is $slightly clean|slightly dirty|completely clean|completely dirty$ right now, and the new bus boy is clearly responsible. He seems very young, this new guy.

Is this table clean now? NO YES

2. This building is $slightly safe|slightly dangerous|completely safe|completely dangerous$ in high winds, at least according to the building inspector. He was here last week.

3. This student is $slightly healthy|slightly sick|completely healthy|completely sick$ all the time, at least according to his own report. But I don't know whether to believe him.

4. This dog is $slightly dry|slightly wet|completely dry|completely wet. right now, from playing near the foul-smelling pond. He's very frisky.

5. This report is $slightly finished|slightly unfinished|completely finished|completely unfinished$

right now, which is the clearly announced deadline for it. The supervisor will want to see it right away.

6. This juice is $slightly pure|slightly impure|completely pure|completely impure$ according to the tests, and the nutritionist doesn't recommend it. I guess it has a high glucose content.

7. This paragraph is $slightly clear|slightly unclear|completely clear|completely unclear$ at the end, but we don't really know who wrote it. The teacher will have to figure it out.

8. This stone is $slightly smooth|slightly rough|completely smooth|completely rough$ on one side, but the kids will use it for their project anyway. They're not exactly perfectionists.

9. This relic is $a little whole|a little cracked|completely whole|completely cracked$ in the photograph, so it's not representative of the collection. It shouldn't be used on the cover of the brochure.

10. This dog is $a little satiated|a little hungry|completely satiated|completely hungry$ right now, so let's wait to play any games with him. Let's play frisbee instead.

11. This road is $a little straight|a little curved|completely straight|completely curved$ in Ohio, so it's more than just a bit tedious to drive. Let's take turns driving.

12. This door is $a little closed|a little open|completely closed|completely open$ right now, so Professor Smith may still be in a meeting. He doesn't like to be interrupted.

13. These dentures are $a little perfect|a little imperfect|completely perfect|completely imperfect$ in terms of the fit, thought the hygienist. She didn't say anything though.

14. This reporter is $a little truthful|a little untruthful|completely truthful|completely untruthful$ in his self presentation, according to the editor. We'll see if that becomes a problem.

15. This guide is $a little certain|a little uncertain|completely certain|completely uncertain$

today, but almost all the other guides are not. Let's radio for another opinion.

16. This caterpillar is $a little dead|a little alive|completely dead|completely alive$ in the documentary, according to William's report. We don't know why.

17. This singer is $slightly secure|slightly insecure|completely secure|completely insecure$ on stage, but she has an absolutely amazing voice. She'll probably go far.

18. This woman's face is $slightly unwrinkled|slightly wrinkled|completely unwrinkled|completely wrinkled$ despite her being 70, at least according to her daughter. I wish my face looked as good.

19. This house was $slightly intact|slightly damaged|completely intact|completely damaged$ after the flood, but now who knows what condition it's in. The inspector won't come for several weeks.

20. This drain pipe was $slightly stopped up|slightly opened up|completely stopped up|completely opened up$ yesterday, but I don't know how it is now. The plumber came, but he quickly left again.

Appendix 4. Materials for Experiment 4. Options are separated by |. Sample interpretation options shown for item 1.

1. The plates were mostly clean|dirty.

What did that mean? 1 = Most of the plates were clean|dirty; 2 = The plates were clean|dirty to a fairly large degree

2. The students were mostly healthy|sick.

3. The guests were mostly satiated|hungry.

4. The bicycles were mostly dry|wet.

5. The relics were mostly whole|cracked.

6. The manuscripts were mostly intact|damaged.

7. The neighborhoods were mostly safe|dangerous.

8. The reports were mostly complete|incomplete.

9. The paragraphs were mostly clear|unclear.

10. The stones were mostly smooth|rough.

11. The eye glasses were mostly perfect|imperfect.

12. The reporters were mostly truthful|untruthful.

Table 1

Mean Proportions of "Acceptable" Choices and Clause 2 Reading Times, Experiment 1

Condition	Pr "Acceptable"	Reaction Time, ms
Maximum standard (5a)	.840	2208
slightly/Maximum standard (5b)	.570	2726
Minimum standard (5c)	.852	1997
slightly/Minimum standard (5d)	.825	2291

Table 2

Mean Proportions of "Acceptable" Choices and Clause 2 Reading Times, Experiment 2

Condition	Pr "Acceptable"	Reading Times, ms	
		Region 1	Region 2
Maximum standard (dry...wetter) (8a)	.322	2322	3423
Minimum standard (wet...drier) (8b)	.468	2443	3512

Table 3

Reading Time Measures, by Region of Sentence, Experiment 3

Condition	Region 1 (Initial)	Region 2 (adv + adj)	Region 3 (next phrase)	Region 4 rest of line
	First Pass Reading Times, ms			
Slightly Maximum standard (9a)	616	502	477	824
Slightly Minimum standard (9b)	588	520	467	802
Completely Maximum standard (9a)	560	523	449	835
Completely Minimum standard (9b)	585	534	457	806
	Second Pass Reading Times, ms			
Slightly Maximum standard (9a)	67	193	82	66
Slightly Minimum standard (9b)	40	109	35	26
Completely Maximum standard (9a)	44	103	50	39
Completely Minimum standard (9b)	34	77	41	31
	Total Reading Times, ms			
Slightly Maximum standard (9a)	678	691	555	897
Slightly Minimum standard (9b)	617	625	500	835
Completely Maximum standard (9a)	595	621	497	878
Completely Minimum standard (9b)	610	601	493	845