Evaluativity across adjective and construction types: An experimental study

Adrian Brasoveanu[1] & Jessica Rett[2]
[1]UCSC & [2]UCLA

July 1, 2016

Abstract

An adjectival construction is evaluative if and only if it conveys that the property associated with the adjective exceeds a relevant threshold. The questions of which adjectival constructions are evaluative and why have formed the foundation for semantic theories of these constructions and of adjectives themselves (Klein 1980, von Stechow 1984), although it's been alleged that these theories are based on an incomplete picture of the phenomenon of evaluativity (Bierwisch 1989, Rett 2008a).

We present the first experimental tests of the scope and nature of evaluativity across adjectival constructions and adjective types. These studies confim that evaluativity is conditioned by adjective type (relative or absolute, Kennedy and McNally 2005) and is not restricted to the positive construction. However, they also show several new and surprising aspects of evaluativity: that it is perhaps better characterized as a gradable property than a binary one; that the ways in which relative and absolute adjectives differ in their evaluativity vary across construction; and that, contrary to standard intuitions, subjects are willing to attribute evaluativity to the subject position of comparative constructions like *Sue is taller than Bill*. We show that this last particularly surprising result reveals a lot about how subjects interpret contextually-sensitive constructions, and we discuss its consequences for experimental studies and semantic theories of adjectival evaluativity as well as context-sensitive phenomena more generally.

Acknowledgments We thank three anonymous *Journal of Linguistics* reviewers for their comments, as well as Sam Cumming, Donka Farkas, Chris Kennedy, Megan Moodie, Floris Van Vugt, Kaeli Ward, and Lauren Winans for discussion and advice. The usual disclaimers apply.

1 Introduction

Example (1-a) exemplifies a positive construction: a construction containing an unmodified gradable adjective. Example (1-b) exemplifies a comparative.

(1) a. John is tall.
b. John is taller than Sue.

There is an important difference between the two constructions, one which has been central to a long-standing debate about the semantics of gradable adjectives: while (1-a) entails that John counts as significantly tall in the context of interpretation, (1-b) does not. In other words, (1-b) but not (1-a) is acceptable in a context in which John does not qualify as tall.

Intuitively, it is the adjective *tall* in (1-a) that contributes this "taller than average" or "significantly tall" meaning, which we'll refer to as 'evaluativity'.[1] But as Klein (1980) has pointed out, the difference

[1]This term is reminiscent of, but unrelated to, the label for "evaluative adjectives" like *beautiful* and *lazy*.

between (1-a) and (1-b) poses a problem for a compositional semantics of adjectival constructions: if evaluativity is lexically encoded in the adjective *tall*, what suppresses it in the comparative in (1-b)? And if it isn't, what introduces it in the positive construction in (1-a)?

Since Klein's discussion (and preceding work in Bartsch and Vennemann 1972, Kamp 1975, Cresswell 1976), evaluativity remains a central issue for the semantics of adjectival constructions (and, consequently, degree semantics generally), and it remains a central member of the class of context-sensitive phenomena. Evaluativity has been argued to vary across adjectival classes (Rotstein and Winter 2004, Kennedy and McNally 2005); across adjectival constructions (Bierwisch 1989, Rett 2008b); and in how it contributes semantically to a context of utterance (Barker 2002, Rett 2015).

As it stands, the discussion of the distribution and contribution of evaluativity has been largely relegated to the theoretical literature.[2] The goal of this paper is to present some initial forays into an experimental investigation of the distribution and contribution of evaluativity. We take for granted Bierwisch's (1989) claim that evaluativity is not distributionally restricted to the positive construction, and we focus on two other questions: first, how, if at all, do relative and absolute adjectives behave differently across constructions with respect to evaluativity? And second, how, if at all, does context play a role in the distribution of evaluativity across adjectival constructions?

We begin in §2 by reviewing evaluativity generally, as well as the various adjectival classes and constructions included in the experiment. In §3 we present a series of experiments intended to address the first question, about the effect of adjective class on evaluativity. After discussing a particularly surprising result of these experiments (and the history of this result in the psychology literature, in §3.2.4), we present in §4 a third experiment designed to follow up on the first two by further probing the role of context in the distribution of evaluativity (the second question). While our results are in general consistent with previous claims about evaluativity, our results suggest that the phenomenon is idealized in the theoretical literature: evaluativity arises in some circumstances as a pragmatic inference, and it seems more gradable than categorical.

2 Evaluativity: relevant data and current accounts

A gradable adjectival construction is evaluative if and only if it conveys that the property associated with the adjective is instantiated to a degree above a particular standard or threshold. In the case of relative adjectives, this amounts to conveying that the degree is significantly high (see §2.2 for a discussion of evaluativity in other types of gradable adjectives). The positive construction in (2-a) – characterized by its lack of any overt measure phrase (e.g. *5ft*), degree modifier (e.g. *very*), or degree quantifier (e.g. *more*) – is evaluative, while the measure phrase (MP) construction in (2-b) and the comparative in (2-c) are not (Cresswell 1976, Klein 1980, von Stechow 1984, Seuren 1984, Bierwisch 1989).

(2) a. John is tall.
b. John is 5ft tall.
c. John is taller than Sue.

In particular, (2-a) entails that John's height exceeds a salient standard of tallness; the sentence is, for instance, incompatible with the claim that John is short. This is such an intrinsic property of positive constructions that the observation seems trivial, until the positive construction is considered alongside other adjectival constructions like the ones in (2-b) and (2-c). These are not evaluative; they do not entail

[2] Although see Sassoon and Zevakhina (2012) for an investigation of the evaluativity of constructions containing degree modifiers like *completely* and *slightly* and the poster by Kim et al. (2013) for an early report on investigations more closely related to the ones reported in the present paper. Paradis and Willners (2006), Zevakhina and Geurts (2011), Doran et al. (2009) present experiments which test the relationship between antonyms, and Barner and Snedeker (2008) and McNabb (2012), Solt and Gotzner (2012) report experiments on children and adults, respectively, in terms of how the standard invoked in evaluativity is calculated.

that John (or Sue) is tall. Both are in principle (depending on the context) compatible with the claim that John (and Sue) is not tall, as (3) demonstrates.

(3) a. John is tall... *(although/in fact) he's not tall.
b. John is 5ft tall... (although/in fact) he's not tall.
c. John is taller than Sue... (although/in fact) he/she is not tall.

Historically, this contrast in evaluativity has been the central worry for compositional semantic accounts of adjectives (Cresswell 1976, Klein 1980, 1982, von Stechow 1984, van Rooij 2008, Rett 2015). It is extremely tempting to tie the evaluativity of (2-a) to the adjective *tall*, but this move seems at odds with the fact that (2-b), which is formed by adding a measure phrase to this same string of morphemes, is not evaluative.

Roughly two distinct treatments of evaluativity have arisen as a consequence: (*i*) those that encode evaluativity in gradable adjectives and argue that this meaning is compositionally 'suppressed' in MP constructions and comparatives (Klein 1980, 1982, Neeleman et al. 2004, Doetjes et al. 2009, Burnett 2012, Cobreros et al. 2012); (*ii*) and those that do not encode evaluativity in gradable adjectives and argue that it is contributed, where available, by a null morpheme like POS (Bartsch and Vennemann 1972, Cresswell 1976, Kennedy 1999; see Rett 2008a,b for the proposal of a distinct null morpheme EVAL).[3,4]

These two different approaches to evaluativity reflect very different perspectives on larger issues: whether a semantic theory that includes null morphemes is strictly speaking a compositional one; whether the semantics of gradable adjectives and the constructions they occur in warrant adding an additional primitive – degrees – to the ontology, etc. However, a direct comparison of these theories is complicated, as they rely on the picture of evaluativity illustrated in (2), which is not in fact representative of the phenomenon. In the rest of this section, we discuss two additional considerations in the distribution of evaluativity: (*i*) its relatively wider distribution across adjectival constructions; and (*ii*) its relatively complicated distribution across adjective classes.

2.1 Evaluativity across adjectival constructions

As (2-a) and (3-a) indicate, evaluativity is an intrinsic property of positive constructions. As a result, we can test for the presence of evaluativity in more complex adjectival constructions by testing whether that construction entails its corresponding positive construction.[5] This has been labeled 'the Bierwisch Test' in Rett (2008b), based on its use in Bierwisch (1989). This test is illustrated in (4).

[3]There are a few other approaches to evaluativity. Breakstone (2012) encodes evaluativity in adjectival meanings and uses a null morpheme 'SSM' to eliminate it when necessary; Rett (2015) argues that evaluativity arises when it does as a conversational implicature. See Rett 2015 for a more comprehensive overview.

[4]In particular, POS accounts typically assume that gradable adjectives denote relations between an individual and a degree of measure, as in (i) below (Cresswell 1976, Hellan 1981, von Stechow 1984). In MP constructions like (3-b), the MP is assumed to saturate the adjective's degree argument.

(i) $[\![\mathsf{tall}]\!] = \lambda x_e \lambda d_d.\ \mathsf{tall}(x, d)$

From the perspective of (i), positive constructions like (3-a) lack any overt morpheme to saturate or bind the adjective's degree argument; they also lack a clear source for the evaluative meaning. Cresswell proposed POS to solve both of these problems simultaneously: it contributes evaluativity to the compositional meaning of the construction; and it binds the adjective's degree argument, allowing the positive construction to denote a proposition. A typical formulation is in (ii-a), from Kennedy and McNally (2005), where G ranges over gradable adjectives (type $\langle e, \langle d, t\rangle\rangle$) and C is a pragmatic variable whose value is the context of evaluation. The resulting interpretation for the positive construction in (3-a) is provided in (ii-b).

(ii) a. $\text{POS} \rightsquigarrow \lambda G_{\langle e,\langle d,t\rangle\rangle} \lambda x_e.\ \exists d[\mathsf{standard}(G, d, C) \wedge G(x, d)]$
b. $[\![\text{John is POS tall}]\!] = \exists d[\mathsf{standard}(\mathsf{tall}, d, C) \wedge \mathsf{tall}(\mathsf{john}, d)]$

[5]Of course, because evaluativity is a context-sensitive phenomenon, we have in mind the notion of entailment between constructions *holding fixed the context of evaluation*. This is distinct from the notion of logical entailment, which is not context-sensitive.

(4) a. John is tall. $\rightarrow$ John is tall.
b. John is 5ft tall. $\not\rightarrow$ John is tall.
c. John is taller than Sue. $\not\rightarrow$ John is tall.

The positive construction *John is tall* – the only evaluative construction of the three – is entailed, of course, by itself in (4-a). But it is not entailed by either the MP construction in (4-b) or the comparative in (4-c): we cannot infer from a report of John's height that he is tall because whether or not 5ft counts as tall is context-dependent. We can similarly not infer from John and Sue's relative heights that John is tall because it is possible for John to be taller than Sue in a context in which they both count as short.

A second test, the Negative Antonym Entailment Test, is illustrated in (3). It mimics the Bierwisch Test while avoiding the issue of self-entailment in (4-a): non-evaluative constructions are compatible with negated versions of their positive-construction counterparts, while evaluative constructions are not. We will adopt this family of tests for evaluativity: a construction is evaluative iff it entails its positive-construction counterpart (or, alternatively, iff it is incompatible with the negation of its positive-construction counterpart).[6]

The two most influential theories of evaluativity – the POS account proposed in Bartsch and Vennemann (1972) and the degree-free semantic account proposed in Klein (1980, 1982) – take for granted that the data in (3) and (4) are representative of the distribution of evaluativity. In particular, their accounts of evaluativity are based on the assumption that evaluativity is in complementary distribution with overt degree morphology, which we use as an umbrella term for MPs like *5ft* (usually taken to be degree-denoting referential terms) and comparative morphemes like *more/-er* (usually taken to denote degree quantifiers). These theories and their recent adaptations were therefore built to make the following prediction: evaluativity arises only when the gradable adjective's degree argument is not valued by an MP or bound by a degree quantifier. The POS account detailed in footnote 4 does this by encoding evaluativity in a degree quantifier: POS can only apply when the adjective's degree argument has not been valued by an MP or bound by another, overt degree quantifier. If it has been, POS cannot apply to contribute evaluativity.

There is, however, reason to doubt that evaluativity occurs only in the absence of overt degree morphology. Bierwisch (1989) and Rett (2008a,b) argue that evaluativity is a property of some (but not all) degree constructions with negative antonyms. These are shown in (5), in contrast to the non-evaluative positive-antonym constructions in (6).

(5) a. John is as short as Sue. $\rightarrow$ John/Sue is short.
b. Sue knows how short John is. $\rightarrow$ John is short.
c. John is this/yea short. $\rightarrow$ John is short.

(6) a. John is as tall as Sue. $\not\rightarrow$ John/Sue is tall.
b. Sue knows how tall John is. $\not\rightarrow$ John is tall.
c. John is this/yea tall. $\not\rightarrow$ John is tall.

The sentences in (5) all involve the negative antonym *short*:[7] (5-a) is an equative – while comparatives are assumed to encode a strict, $>$ ordering on degrees, equatives are assumed to encode a non-strict, $\geq$ ordering (Horn 1972, Klein 1980, Bierwisch 1989, Bale 2008, Rett 2014); (5-b) is a degree relative – the evaluativity of this clause also carries over to the question *How short is John?*; and (5-c) is a degree

[6] For degree constructions that contain truth-conditional operators, like negation, this test must be modulated. The negated positive construction in (i) does not pass the Bierwisch Test, for obvious reasons:

(i) John is not tall. $\not\rightarrow$ John is tall.

However, the negated positive construction is still considered evaluative. Specifically, the evaluative component of a positive construction seems to semantically scope under negation. *John is not tall* means it's false that John is tall to a significant degree; it doesn't amount to the negation of the weaker proposition 'John has a height'.

[7] For a discussion of what constitutes the negative or marked antonym, see Cruse (1980, 1986), Lehrer (1985), Rett (2015).

demonstrative.

The constructions in (5) are evaluative, while their positive-antonym counterparts are not. This is not the case for MP constructions, which cannot be formed with negative antonyms (cf. **John is 5ft short*, although see Doetjes 2012). It is also not the case for the comparative construction, which seems to be non-evaluative regardless of the antonym involved, as (7) shows.

(7) a. John is taller than Sue. $\not\rightarrow$ John/Sue is tall.
b. John is shorter than Sue. $\not\rightarrow$ John/Sue is short.

Thus, the distribution of evaluativity is more complicated than traditionally assumed: it is not simply restricted to constructions without overt degree morphology. The positive construction is evaluative as a rule, but some constructions with overt degree morphology are also evaluative when they contain negative antonyms.

There are two properties of the evaluative constructions in (5) that will be relevant here. First, while the evaluativity associated with positive constructions seems to be part of the at-issue content of the sentence, the evaluativity associated with the constructions in (5) seems to be presupposed, or at least not-at-issue; the negation in (8) targets the equative relation, not the evaluativity (Potts 2012, Tonhauser et al. 2013).

(8) A: John is as short as his mother.
B: No, he's not, he's taller!
B: #No, he's not, he's tall for his age!

Because the Bierwisch Test treats evaluativity as an entailment of a construction, however, it doesn't distinguish between at-issue and not-at-issue evaluativity. See Rett (2015) for a more extensive discussion and derivation of at-issue and not-at-issue evaluativity.

Second, in evaluative comparison constructions like the equative in (5-a), evaluativity is asymmetrically associated with the 'object' (Rett 2008b, 2015). We use the terms 'subject' and 'object' to refer to the subject of the external clause and the subject of the internal clause in comparison constructions, respectively; for example, in (9), *John* is the subject, and *Sue* the object. This is demonstrated by the standard projection tests in (9) and (10): the negative-antonym equative presupposes that the object, Sue, is short, but not that the subject, John, is short. Rett (2015) analyzes this asymmetry in terms of at-issueness.

(9) John isn't as short as Sue.
a. $\not\rightarrow$ John is short.
b. $\rightarrow$ Sue is short.

(10) If John is as short as Sue, he will not make the team.
a. $\not\rightarrow$ John is short.
b. $\rightarrow$ Sue is short.

In sum, the characterization of the distribution of evaluativity presented in (2) – and, as a result, semantic accounts of evaluativity – are challenged by claims that evaluativity is present in degree constructions that contain overt degree morphology. While some have interpreted this as a motivation to alter the traditional approaches to evaluativity (Rett 2008a,b, 2015, Breakstone 2012, Grano 2012), it is also possible to interpret the difference in how evaluativity is instantiated in positive constructions and equatives to indicate that there are two types of evaluativity, warranting distinct treatments. And while the intuitions discussed here have been taken to be cross-linguistically universal, there is relatively little work done in testing this hypothesis (although Bogal-Allbritten 2015, 2013, Grano 2012, Schwarzschild 2012, Bochnak and Bogal-Allbritten 2014 are notable exceptions).

2.2 The relative/absolute distinction

Another complication to the traditional picture of evaluativity is that the pattern depicted in (2) seems to hold only for relative adjectives, a subclass of gradable adjectives. Other classes of adjectives have been reported to display other evaluativity patterns (Cruse 1976, Yoon 1996, Rotstein and Winter 2004, Kennedy and McNally 2005).

2.2.1 Empirical diagnostics

Adjectives may or may not be gradable. A good test for gradability is whether the adjective can be intensified (11) or can occur in a comparative construction (12) based on its intrinsic scale, i.e. without being coerced onto a scale of temporality or prototypicality, as in (13).

(11) a. *John is very childless.
b. John is very tall.

(12) a. *John is more childless than Sue.
b. John is taller than Sue.

(13) a. Mary is more pregnant than Sue. *comparison of times*
b. Dogs are more mammalian than whales. *comparison of prototypicality*

Within the category of gradable adjectives, antonym pairs can fall into two distinct categories: relative and absolute.[8] This distinction is based on several diagnostics, discussed in Cruse (1976, 1986), Yoon (1996), Rotstein and Winter (2004), Kennedy and McNally (2005), Kennedy (2007) and Burnett (2012); we review only a few of them here.

Following Kennedy (2007) and Syrett et al. (2010), we take the 'definite description test' to diagnose the difference between relative and absolute adjectives (see Figure 1 for a schema of these subclasses). In a context in which there are several glasses, all differing in heights, a definite description formed with a relative adjective, as (14-a), picks out the individual that instantiates the gradable predicate to the highest degree (i.e., the tallest one), *regardless of whether or not the individual counts as tall in the context*. In contrast, a definite description formed with an absolute adjective, as in (14-b), has a referent only in a context in which a glass actually counts as empty. In a context in which there are several glasses, all containing different levels of liquid (but none empty), (14-b) fails to have a referent.

(14) a. Pass me the tall one. *relative*
b. Pass me the empty one. *absolute*

Kennedy's explanation for this contrast is that relative adjectives, but not absolute adjectives, invoke context-dependent standards.

Within the category of absolute adjectives, there are several subtypes, depicted in Figure 1. These subtypes are differentiated in part by their distribution with a certain class of adjectival modifiers. We will focus on *almost* and *slightly* (Rotstein and Winter 2004). In a relative antonym pair, neither antonym can be modified by either of these modifiers:[9]

(15) John is ??almost/??slightly tall/short. *relative*

In contrast, in closed absolute antonym pairs, both antonyms can occur with both modifiers:

(16) a. The window is almost/slightly opaque. *closed absolute*

[8] Paradis (2001) additionally discusses 'extreme' adjectives – like *stupid* or *gorgeous* – which seem to be intrinsically evaluative. We will not discuss these here; see Morzycki (2012) for an in-depth look at extreme adjectives.

[9] This seems to be the received wisdom in the literature, but see Bogal-Allbritten (2015) and references therein for counter-arguments and more discussion.

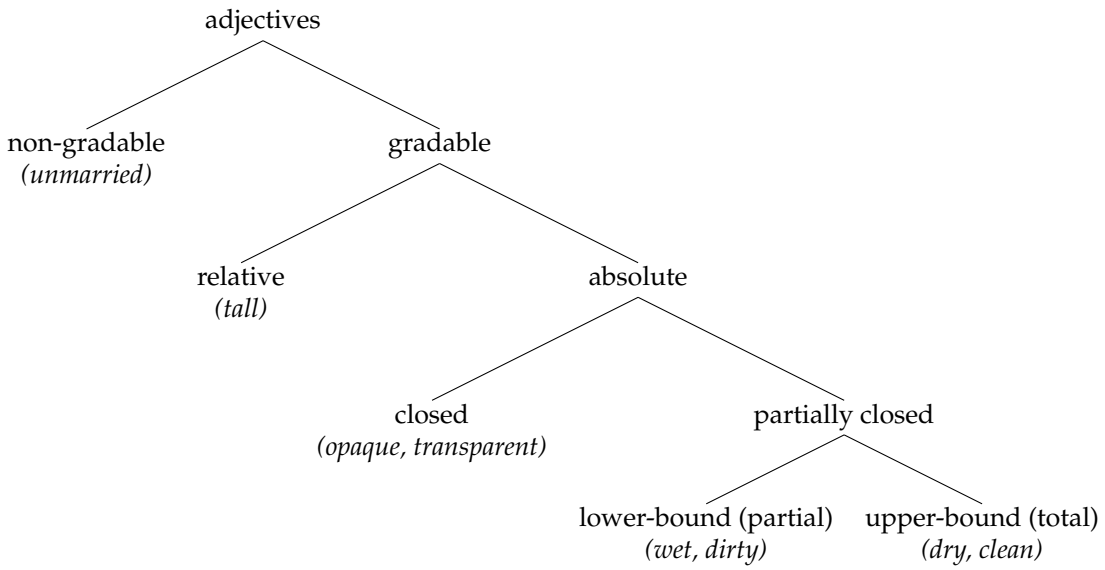

Figure 1: Adjective classification.

b. The window is almost/slightly transparent. *closed absolute*

The partially closed subclass of absolute adjectives differ with respect to which modifier they are compatible with. Total adjectives can be modified by *almost* but not by *slightly*; and partial adjectives can be modified by *slightly* but not by *almost*. (The total/partial distinction is called 'universal/existential' in Kamp and Rossdeutscher 1994, Yoon 1996, Rotstein and Winter 2004.)

(17) a. This towel is ??almost/slightly wet. *partial*
b. This towel is almost/??slightly dry. *total*

(18) a. The towel is ??almost/slightly dirty. *partial*
b. The towel is almost/??slightly clean. *total*

Beginning with work in Kennedy and McNally (2005), the differences between these classes have been attributed to differences in the structure of the scales with which these adjectives are associated. They propose that absolute adjectives lexicalize bounds on their associated scales, while relative adjectives do not. The difference between total and partial adjectives is thus claimed to be that the former lexicalize an upper bound on the scale (a maximum endpoint), while the latter lexicalize a lower bound on the scale (a minimum endpoint). This claim is bolstered in part by independent evidence that *almost* selects for scales with upper bounds, while *slightly* selects for scales with lower bounds.

The closed absolute adjectives in (16) are thought to lexicalize both an upper and a lower bound, hence their compatibility with both modifiers. In contrast, because relative adjectives do not encode any bound, they cannot occur with these modifiers, as demonstrated in (15).

2.2.2 Evaluativity and the absolute/relative distinction

Kennedy (2007) argues that these lexicalized bounded scales have specific consequences for the phenomenon of evaluativity. In particular, he proposes that if an adjective lexicalizes a scale with a bound, that bound is co-opted as an evaluative standard. If it doesn't, context values the standard. His Principle of Interpretive Economy (Kennedy 2007: 36) is reproduced in (19).

(19) KENNEDY'S PRINCIPLE OF INTERPRETIVE ECONOMY
Maximize the contribution of the conventional meanings of elements of a sentence to the com-

putation of its truth conditions.

Because relative adjectives contain no conventional meaning that could correspond to a standard, the standards invoked by relative adjectives (but not absolute adjectives) vary from context to context. This is illustrated in Figure 2.

If this is right, then there are clear consequences for the study of evaluativity. The distribution of evaluativity reported in §2.1 characterizes the behavior of relative adjectives, while the Kennedy/McNally theory predicts that absolute adjectives have a different relationship to the standards by which evaluativity is calculated. In the case of relative adjectives, an individual can be on the scale (say, of tallness) without being above the standard (which could be, say, 6ft tall). However, the Kennedy/McNally theory predicts that partial adjectives lexicalize a scale with a lower bound, and that this lower bound is co-opted as a standard for the purposes of evaluativity. The result is the prediction that any construction containing a partial adjective will count as evaluative for at least one argument, because any individual that falls on the scale of a partial adjective (e.g. *dirty*) will necessarily be an individual that exceeds the standard of dirtiness (i.e., the scale's lower bound). In a comparative formed from a partial adjective, *x is Adj-er than y*, the subject argument is predicted to be evaluative.

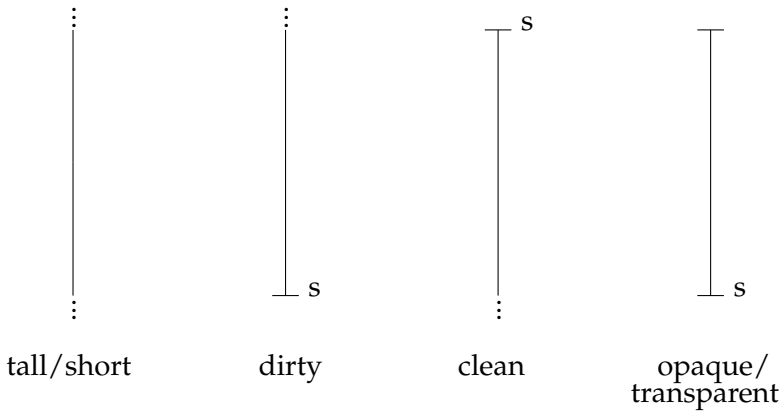

Figure 2: Adjective classes and scale bounds

In contrast, total adjectives lexicalize a scale with an upper bound, and this bound (given the Principle of Interpretive Economy) is co-opted as a standard for the calculation of evaluativity. As a consequence, any individual on a scale associated with a total adjective is located at or below the standard. The result is that for any construction containing an total adjective that differentiates between more than one point on the scale will be non-evaluative for at least one argument.

Closed absolute adjectives like *opaque* and *transparent* lexicalize a scale with both an upper and lower bound. Kennedy's Principle of Interpretive Economy is silent on which of two bounds should be co-opted for the purpose of evaluativity, and it seems as though the class of closed adjectives is not uniform in this respect. Some closed adjectives (e.g. *opaque/transparent*) have been characterized as having variable interpretations, in which either bound can function as a standard, while others (e.g. *full/empty*) seem to only use the maximum standard (Kennedy 1999).

These predictions are largely substantiated by experimental work; see Kennedy and Levin (2008), Morzycki (2009), Syrett et al. (2010), McNally (2011), Toledo and Sassoon (2011) for more evidence in support of the Kennedy/McNally theory generally. In particular, while relative adjectives display the evaluativity pattern presented in §2.1, partial absolute adjectives are evaluative across constructions, and total absolute adjectives are only evaluative in the positive construction, as illustrated in (20)-(23)[10] and summarized in Table 1. And there has been additional experimental evidence that either bound can function as a standard in degree constructions formed from some closed absolute adjectives Sassoon (2012a).

[10]These only test for evaluativity in the object position, given the asymmetry reported in (9) and (10).

However these judgments, in contrast to the relative adjective judgments, have been reported to be weaker and more controversial; see in particular Toledo and Sassoon (2011), but also Sassoon (2012b), for claims that the relative adjective data are more variable than reported here.[11]

(20) TOTAL ABSOLUTE

a. This shirt is cleaner than these jeans. COMP
↛ These jeans are clean.

b. This shirt is as clean as these jeans. EQ
↛ These jeans are clean.

(21) PARTIAL ABSOLUTE

a. This shirt is dirtier than these jeans. COMP
→ These jeans are dirty.

b. This shirt is as dirty as these jeans. EQ
→ These jeans are dirty.

(22) CLOSED ABSOLUTE TYPE A

a. The front window is more opaque/transparent than the back window. COMP
→ The back window is opaque/transparent.

b. The front window is as opaque/transparent as the back window. EQ
→ The back window is opaque/transparent.

(23) CLOSED ABSOLUTE TYPE B

a. The front display window is more full/empty than the back display window. COMP
→ The back display window is full/empty.

b. The front display window is as full/empty as the back display window. EQ
→ The back display window is full/empty.

	RELATIVE		ABSOLUTE		
	POSITIVE	NEGATIVE	CLOSED	TOTAL	PARTIAL
POSITIVES	yes	yes	yes/no	yes	yes
EQ(UATIVES)	**no**	yes	yes/no	**no**	yes
COMP(ARATIVES)	**no**	**no**	yes/no	**no**	yes

Table 1: Distribution of evaluativity

2.3 Theoretical summary

The first goal of the present study is to provide an experimental test for the predictions of the Kennedy/McNally theory, which characterizes different adjective classes in terms of their scale structure, and in particular in terms of which (if any) scalar bounds they lexicalize. If the difference between relative and absolute adjectives is really a difference in scale structure – and, in turn, a difference in the availability of a contextually stable standard of comparison – then we predict the two classes of adjectives will behave differently in the distribution of evaluativity across degree constructions and across contexts.

[11] A reviewer worries that the Kennedy/McNally treatment doesn't make claims about evaluativity in terms of logical entailment but rather in terms of (pragmatic) implicature, based on the assumption that Kennedy's (2007) Principle of Interpretive Economy is a violable constraint (like Grice's maxims) rather than an inviolable one. Unfortunately, Kennedy does not formalize this principle in any way, so the theory's precise predictions are unclear. We assume here that the principle, while pragmatic, holds without exception, given the absence of any evidence that Kennedy intended it to be violable or cancelable (or that it must be violable or cancelable). It is, however, possible that some formal implementation of the Principle of Interpretive Economy would predict that evaluativity arises in these cases only as a strong implicature, rather than an entailment.

The pattern of evaluativity predicted by the Kennedy/McNally theory is generally supported by the introspective judgments provided in the theoretical literature (e.g. Rett 2008a,b, Morzycki 2009, Syrett et al. 2010, McNally 2011), but the judgments are regularly reported as relatively weak or variable (see Rett 2015 §5.4 for an in-depth discussion of this variation).

There are several possible explanations for the variation in these judgments. First, as discussed in §2.1, the standard tests for evaluativity are somewhat flawed. The Bierwisch Test and the Negative-Antonym Entailment Test only detect evaluativity in clauses that are not embedded under truth-conditional operators. The Bierwisch Test has an additional flaw in that it can't be applied equally across degree constructions: for positive constructions, the Bierwisch Test amounts to a tautology, a test of whether e.g. the sentence *John is tall* entails the sentence *John is tall*. In other words: the reported judgments of evaluativity could be variable because we're not testing them properly (or consistently, etc.).

Second, speaker intuitions about evaluativity could be variable because evaluativity is context-sensitive. This is certainly true for the evaluativity associated with relative adjectives, but also arguably true for absolute adjectives. Kennedy's Principle of Interpretive Economy exhorts the speaker to maximize the conventional meanings of a sentence, but an absolute adjective's lexicalized standard will no doubt vary across comparison classes: as Whorf pointed out (1956: 135), standards of emptiness seem to vary between bottles of beer and drums of gasoline. Depending on the formal status of Kennedy's principle, it might be overruled by other considerations in some contexts of utterance but not others.

Third, speaker intuitions about evaluativity could be variable because they are by nature variable; Rett (2015) proposes that evaluativity arises as a conversational implicature, and suggests that variation in evaluativity across antonym pairs could be attributed to the different extents to which evaluativity as an implicature is conventionalized for those antonym pairs based on potential differences in frequency, breadth of meaning, and in the relationship between antonyms in a pair (Lehrer 1985).

From this perspective, the phenomenon of evaluativity is a compelling area for experimental investigation. Controlled experiments have the potential to contribute consistency and subtlety to the debates outlined above. But because evaluativity is relatively uncontroversial as a general phenomenon, it is also tempting grounds on which to explore experimental methods of context-sensitive phenomena.

We begin our experimental tests of evaluativity in §3 by addressing the extent to which relative and absolute adjectives differ in their evaluativity across degree constructions. The results we report largely confirm the predictions made by the Kennedy/McNally theory, with the exception of one particularly surprising observation: contrary to reports in the linguistic literature, the data suggest that comparatives are evaluative. We further investigate this result in a third experiment, reported in §4, which we argue illuminates an important difference in context-sensitivity between informal introspective judgments and at least some experimental methodologies.

3 Evaluativity and adjective class

The goal of these experiments is to test the distribution of evaluativity across degree constructions and adjective classes. The experiments share a common methodology: a Bierwisch Test couched in what we call a "justifiable entailment task," similar in general respects to the "coherence acceptability task" in Experiment 3 of Cummins and Katsos (2010) (see also Katsos 2008). Participants were presented with a series of deductions that a Police Chief makes based on one-sentence reports from his Detective. The participants were asked to evaluate how justified the Chief's conclusions were on a 5-point Likert scale: -2 (not justified at all), -1 (somewhat unjustified), 0 (neither justified nor unjustified), 1 (somewhat justified), 2 (strongly justified).

An example stimulus is provided in Table 2 below:

The Detective reported to the Police Chief: "Maria is as short as Sophie."
The Chief concluded from this that Sophie is short.

Based on the Detective's report, how justified was the Chief in drawing that conclusion?

−2	−1	0	1	2
not justified at all				strongly justified

Table 2: Example stimulus

All the stimuli (items or fillers) had the same format, the only parts that changed were the Detective's quoted report *The Detective reported to the Police Chief: "…"* and the Chief's conclusion in the subordinate clause *The Chief concluded from this that …*. As we will discuss in detail below, the target sentence (in the Detective's report) changed according to degree construction and adjective class.

This methodology uses speakers' judgments about whether a conclusion is justified as a proxy for entailment in a context of evaluation, which is a discourse variant of the Bierwisch Test.[12] As a consequence of this difference, the judgments tested by this experimental design are more context-sensitive than the results of the Bierwisch Test. Presenting the data in a conversation might mean that subjects are more likely to interpret the utterance with some conversational implicature; incorporating a second agent might signal to subjects that two different standards could be at play. In §3.2 we will argue that our results suggest that subjects are indeed enriching these sentences with context in a way that the Bierwisch Test doesn't predict.

A pilot study allowed us to test this methodological approach and confirm that it was a reliable indication of evaluativity. The results of this pilot were replicated by our first main experiment, discussed immediately below, so we will not present the pilot here. See Appendix A for those data and results.

3.1 Method and results

Before turning to the discussion of the experimental method and the results, we briefly recapitulate the predictions made by the previous linguistic literature in terms of our expectations for the experiment at hand: (i) we expect the positive construction to be evaluative *tout court* – in particular, we expect no difference in evaluativity across adjective classes; (ii) we expect to see a difference in evaluativity between positive and negative polarity adjectives in equatives, but not in positive constructions or in comparatives (in either subject or object position); and (iii) we expect to see some differences in evaluativity between relative and some absolute adjectives for in comparatives, as detailed in Table 1.

3.1.1 Method

The main goal of Experiments 1 and 2 was to test the difference in evaluativity across the relative and absolute adjectival classes, and with respect to different construction types and adjective polarities. Adjective class was a between-subject variable, i.e., Experiment 1 tested relative adjectives exclusively, while Experiment 2 tested only absolute adjectives. We will continue to distinguish between these two experiments in the discussion of the method, participants and data summaries. When we turn to the statistical analysis however, we will pool together the data from both experiments and analyze it as one data set. As already indicated, we distinguish between the data collected in Experiment 1 vs. 2 in the statistical analysis by adding an additional factor for adjective class (relative vs. absolute adjectives).

Construction type and adjectival polarity were within-subject variables – all combinations were tested in each of the two experiments. The reason for a between-subjects design for adjective class is

[12] In this respect, it is similar to the methodology used in Cummins and Katsos (2010), but it differs in that the relevant conclusion is made by one agent on the basis of another agent's claim.

that the experiment (including practice items and a reasonable number of fillers) would have been unfeasibly long.

	Experiment 1: rel. adj.		**Experiment 2: abs. adj.**	
	pos. pol.	*neg. pol*	*pos. pol.*	*neg. pol.*
1	tall	short	complete	incomplete
2	dark	light	full	empty
3	expensive	inexpensive	opaque	transparent
4	fast	slow	open	closed
5	hard	easy	perfect	imperfect
6	heavy	light	visible	invisible
7	strong	weak	straight	bent
8	wide	narrow	awake	asleep

Table 3: Experiments 1 and 2: Antonym pairs

Experiment 1 focused on 8 adjective pairs from the relative-adjective class, while Experiment 2 focused on 8 adjective pairs from the absolute-adjective class. These are listed in Table 3 according to their adjectival class.

The absolute adjective class contains both closed absolute adjectives and total/partial antonym pairs. The first within-subject condition was adjectival polarity. In both adjectival classes, polarity was determined based on the markedness tests developed in Cruse (1976, 1986), Lehrer (1985): the negative antonym in a pair is the more morphologically complex member, and the member with the least natural nominalized form, etc. (cf. *openness* vs. **closedness*). For the two total/partial antonym pairs in the absolute condition, the partial antonym was also the negative antonym (cf. *straightness* vs. **bentness*, and *awakeness* vs. **asleepness*).

The already large number of cells in the present experiments prevented us from investigating the specific predictions made in the Kennedy/McNally theory with respect to the difference in evaluativity between total and partial adjectives. But we investigated this issue in the pilot experiment discussed in Appendix §A. In the pilot, there was no significant difference generally (i.e., across all constructions) between positive and negative antonyms in the relative and closed absolute classes, but there was a significant difference between total and partial antonyms. This difference is still monitored in the present experiment using polarity. In general, in terms of evaluativity across constructions, total adjectives behaved in the pilot like relative adjectives, while partial adjectives behaved in the pilot like closed absolute adjectives.

The second within-subjects condition was construction type. We tested for evaluativity in three different constructions: the positive construction, the comparative, and the equative, exemplified in (24).

(24) a. `pos`: *Maria is short*
b. `subjComp/objComp`: *Maria is shorter than Sophie*
c. `subjEq/objComp`: *Maria is as short as Sophie*

However, for reasons discussed in §2.1, we are interested in testing for evaluativity in both clauses of the equative and comparison constructions, so the construction condition has five sub-conditions, illustrated in Table 4.

condition	target sentence	test sentence
`pos`	*Maria is short*	*Maria is short*
`subjComp`	*Maria is shorter than Sophie*	*Maria is short*
`objComp`	*Maria is shorter than Sophie*	*Sophie is short*
`subEq`	*Maria is as short as Sophie*	*Maria is short*
`objEq`	*Maria is as short as Sophie*	*Sophie is short*

Table 4: Sub-conditions of construction type

Eight items were constructed for each of the `posPol-negpol` adjective pairs listed in Table 3. That is, Experiment 1 involved a set of 64 items and Experiment 2 involved a distinct set of 64 items. Every item was passed through 10 conditions (2 adjective polarities $\times$ 5 construction types). Every participant saw each of these items exactly once, either with the `posPol` or the `negPol` adjective, and in one of the 5 construction types, for a total of 64 stimuli (excluding fillers). The adjective polarity & construction type combination was randomly selected for every one of the 64 items and rotated for every participant (Latin square design).

The participants were presented with a series of deductions that a Police Chief makes based on one-sentence reports from his Detective. They were asked to evaluate how justified the Chief's conclusions were on the same 5-point Likert scale: -2 (not justified at all), -1 (somewhat unjustified), 0 (neither justified nor unjustified), 1 (somewhat justified), 2 (strongly justified) (see Table 2 above). (25) and (26) contain example stimuli in the `subjComp` condition, (25) from the relative adjective experiment (Experiment 1) and (26) from the absolute adjective experiment (Experiment 2).

(25) a. The Detective reported to the Police Chief: "Picking locks is harder than cracking safes." The Chief concluded from this that picking locks is hard.
b. The Detective reported to the Police Chief: "The city's sidewalk is wider than the town's sidewalk." The Chief concluded from this that the city's sidewalk is wide.
c. The Detective reported to the Police Chief: "Martha is stronger than Bertha." The Chief concluded from this that Martha is strong.

(26) a. The Detective reported to the Police Chief: "Mary's instructions are more complete than Richard's instructions." The Chief concluded from this that Mary's instructions are complete.
b. The Detective reported to the Police Chief: "The writing paper is more opaque than the drawing paper." The Chief concluded from this that the writing paper is opaque.
c. The Detective reported to the Police Chief: "Carol's blemishes are more visible than Henry's blemishes." The Chief concluded from this that Carol's blemishes are visible.

In addition to the 64 experimental stimuli, there were 67 fillers involving modal expressions: the modal verb *must*; the modal verb *might*; the modal adverb *certainly*; the modal adverb *possibly*; the modal/evidential adverb *apparently*; and the modal/evidential adverb *supposedly*. The fillers were meant to elicit ratings across the entire range: possibility modals expressions were meant to elicit low ratings, while the ones with necessity modal expressions were meant to elicit higher ratings.[13] Thus, every partic-

[13]Examples for each type of filler are provided below:

(i) a. *John must like cake* $\Rightarrow$ *John likes cake*
b. *Mary might have brought a dessert to the dinner* $\Rightarrow$ *Mary brought a dessert to the dinner*
c. *Mary certainly visited the National Gallery on Tuesday* $\Rightarrow$ *Mary visited the National Gallery on Tuesday*
d. *Steve possibly owns a boat* $\Rightarrow$ *Steve owns a boat*
e. *Doug apparently sleeps with a teddy bear* $\Rightarrow$ *Doug sleeps with a teddy bear*
f. *George supposedly has a degree in bioengineering* $\Rightarrow$ *George has a degree in bioengineering*

Aggregating over the responses to the fillers in both Experiment 1 and Experiment 2, we see that the ratings spanned the whole

ipant responded to 131 stimuli (64 experimental stimuli + 67 fillers), the order of which was randomized for every participant.

The sets of participants in the relative and absolute conditions (i.e., in Experiments 1 and 2) were disjoint. 45 participants (undergraduate students from UCLA) completed the relative adjective study online on a UCSC-hosted installation of the IBEX platform (http://code.google.com/p/webspr/) for course credit or extra-credit. 64 measurements were collected per participant, for a total of 2880 observations. 50 different participants (undergraduate students from UCLA) completed the absolute adjective study online on a UCSC-hosted installation of the IBEX platform for course credit or extra-credit. 64 measurements were collected per participant, for a total of 3200 observations.

We examined the pattern of responses that the participants gave for the `pos` (positive) construction to determine if participants properly completed the experiment. This construction is the control, as the Bierwisch Test for evaluativity in positive constructions is a test for whether a sentence entails itself, and we expect an overwhelming amount of 2 or 1 responses for it from any given participant. If a participant has a consistent pattern of very low responses for the `pos` construction, e.g., -2, -1 or 0, relative to the majority of the other participants, it is likely that s/he did not complete the experiment properly and we should drop the data for that participant. We also examined the participants for repeated patterns of behavior that could not be detected by examining the consistently low ratings for the `pos` construction. That is, we checked to see if any participants selected e.g. only 2 or 1 throughout the experiment, which would strongly indicate that they did not complete the experiment properly.

As a result, we dropped the data obtained from 3 participants in Experiment 1 and 2 participants in Experiment 2. The final number of participants in Experiment 1 is 42, for a total of 2688 observations, and the final number of participants in Experiment 2 is 48, for a total of 3072 observations. Aggregating over both the relative and the absolute condition (i.e., over both Experiment 1 and 2), the final number of participants is 90 and the total number of observations is 5760.

---

range and had a bias towards the upper middle of the range: **-2** – 340 (5%), **-1** – 891 (14%), **0** – 1201 (19%), **1** – 2355 (37%), and **2** – 1578 (25%). The previous linguistic literature led us to expect most experimental item ratings to be in the lower part of the range, so the experimental task as a whole was fairly balanced and required participants to regularly use the full range of ratings.

3.1.2 Results

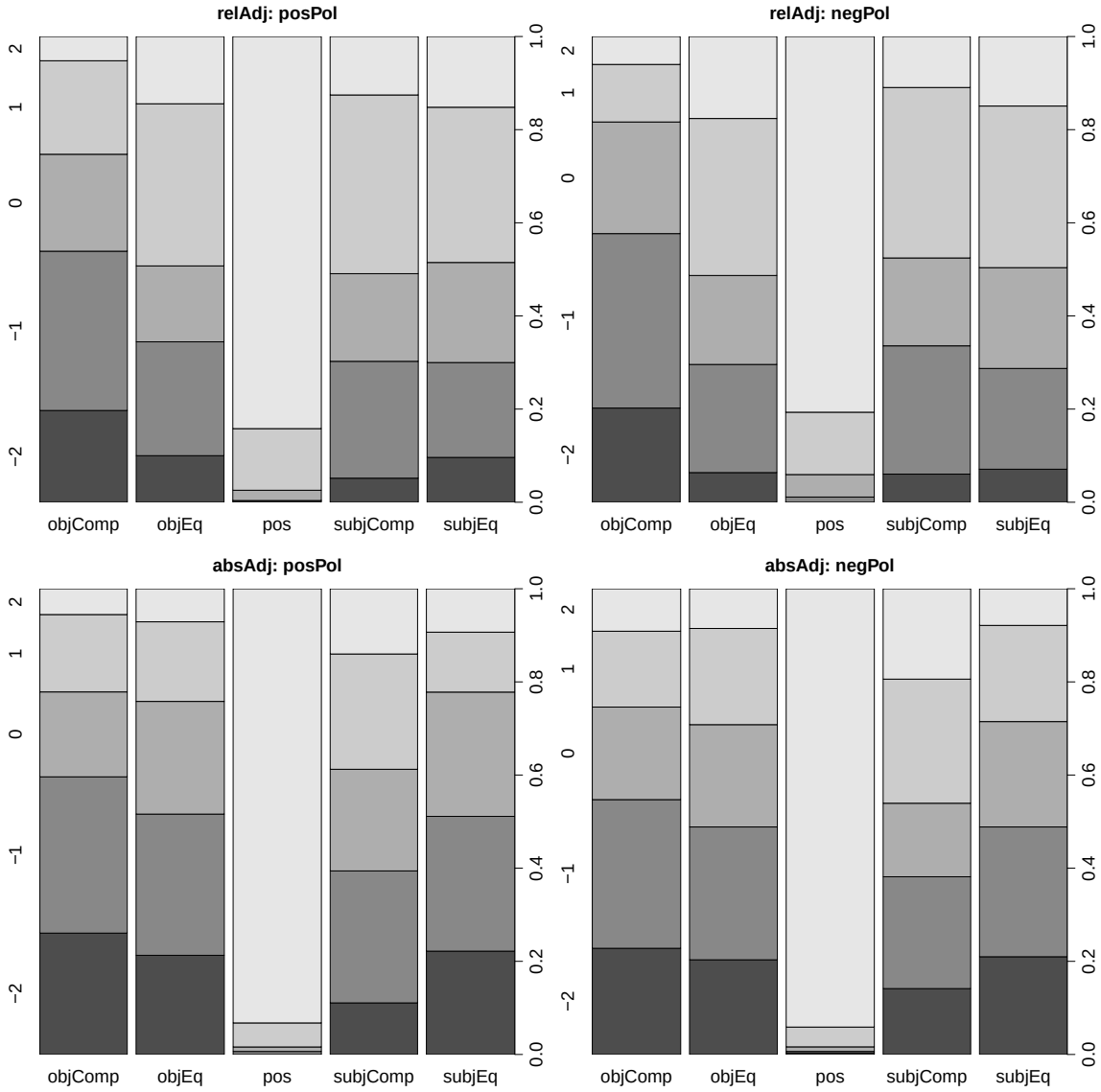

Figure 3: Experiments 1 and 2 – Graphical summaries

Graphical summaries of the data collected in Experiments 1 and 2 are provided in Figure 3.[14] There are 4 plots in that figure. On the first row, we plotted the results for the relative adjective pairs: the plot in the first column from the left corresponds to the positive polarity (`posPol`) adjectives, while the plot in the second column corresponds to the negative polarity (`negPol`) adjectives. The corresponding plots for the absolute adjective pairs are on the second row; once again, `posPol` adjectives are in the first column and `negPol` adjectives in the second. For completeness, the data summaries plotted in Figure 3 are provided in Appendix B.3.

Each plot consists of 5 equal-width, equal-height bars, i.e., bars whose areas are the same. Each bar corresponds to one of the five constructions under study, listed in alphabetical order on the x-axis: `objComp`, `objEq`, `pos`, `subjComp`, `subjEq`. Within each bar, we plot the proportion of 2, 1, 0, -1 and -2 answers in increasingly darker shades of gray. Thus, 2 is nearly white and -2 is nearly black; these

[14] All the data summaries and data analyses in this paper have been obtained / completed using *R* (R Core Team 2013) and the package *Ordinal* (Christensen 2012).

values are specified on the y-axis on the left side of the plot. Within each bar, the area occupied by a particular shade of gray corresponds to the proportion of times the corresponding number has been selected by participants.

For example, if we focus on the top left plot (`relAdj:posPol`), we see that the `pos` construction has received an overwhelming number of 2 (strongly justified) answers. This is indicated by the very big area that is in the lightest shade of gray. The y-axis labels on the right side of each plot help us infer the actual proportion of 2 answers out of the total number of answers given for the `relAdj & posPol & pos` condition: we see that this proportion is more than 0.8, i.e., more than 80%. We also see that about 0.15 of the answers were 1, as indicated by the portion of the bar colored in a slightly darker shade of gray. Finally, there are very few 0, -1 or -2 answers. In contrast, the `objComp` bar of the same plot indicates that very few 2 or 1 answers have been selected for this construction, since most of the bar is occupied by the areas corresponding to 0, -1 and -2.

These plots suggest a set of high-level but richly structured generalizations, taking 'evaluativity' as a proxy for 'acceptability of entailment to the positive construction':

(27) **Generalizations based on the Exp. 1 & 2 data summaries**

a. Adjective polarity:

i. There is no clear difference in evaluativity between `posPol` and `negPol` adjectives within either the `relAdj` class or the `absAdj` class (except for the `objEq` construction);

b. Constructions:

ii. The `pos` construction is clearly the most evaluative: it has received an overwhelming number of 2 answers in all four plots;

iii. but the `relAdj` class is less evaluative than the `absAdj` class in the positive construction;

iv. The `objComp` construction is the least evaluative construction in all four plots while the `subjComp` construction is among the most evaluative;

v. The `objEq` and `subjEq` constructions are basically identical, both within the `relAdj` class and within the `absAdj` class; but the `relAdj` equative constructions are clearly more evaluative than the `absAdj` equative constructions; this is the inverse of the pattern we observed for `pos` constructions;

c. Adjective class:

vi. whether the adjectives in the `relAdj` class are more or less evaluative than the ones in the `absAdj` class depends on construction type: the `relAdj` class is less evaluative than the `absAdj` class for `pos` constructions, more evaluative for `eq` constructions, and the two adjective classes exhibit about the same evaluativity level for `comp` constructions.

These generalizations are all underpinned by significant effects, as the data analysis in the immediately following subsection shows. We return to a discussion of these generalizations and their theoretical relevance in §3.2.

3.1.3 Statistical analysis: Ordinal probit regression models

The data are not continuous but ordered categorical, i.e., the possible answers -2, -1, 0, 1 and 2 are merely ordered labels rather than actual numbers – in much the same way that the grades *A*, *B*, *C*, *D* and *F* in the US grading system, or the different levels on a pain or happiness scale, are merely ordered labels.

Our task is to find out whether different adjective classes, different adjective polarities and different constructions are associated with different levels of evaluativity, while taking seriously the ordered categorical nature of our evaluativity data. To develop a better intuitive understanding of the question we are asking and the way we will quantify it, let us restrict our attention to relative adjectives (`relAdj`) in the positive construction (`pos`), and simply compare the `posPol` and `negPol` conditions, i.e., the positive

polarity and the negative polarity adjectives. This is like trying to find out whether the level of scholarly ability of a group of Ohio students, for instance, is 'significantly' different from the corresponding level of scholarly ability of a group of Michigan students as measured by their letter grades on the same standardized test.

One of the standard ways of quantifying the difference between these hypothetical Michigan and Ohio students while being faithful to the categorical nature of the letter grades is to say that the grades give us an imperfect image of the underlying, unobserved continuous scholarly ability of the two groups of students. This scholarly ability, just like weight and height, is assumed to have a normal (Gaussian) distribution. To answer the question of whether Ohio and Michigan students differ in ability, then, we would use their respective grades to estimate the mean/center of the normal distributions underlying their respective scholarly abilities. If the estimated means of the normal distributions are very close to each other, we will conclude that the two groups of students are most likely identical in scholarly ability. If the estimated means are clearly apart, we will conclude that the two groups have different scholarly ability. This type of statistical model is called an ordinal probit model: 'ordinal' because we deal with ordered categorical data and 'probit' because we assume that the underlying continuous ability has a normal (Gaussian) distribution.

We apply the same statistical model to the results of Experiments 1 and 2, for the same reasons. The estimates are listed in Appendix B.4 and plotted below in Figure 4. Each plot compares two curves corresponding to the probability densities of two different factors. The comparison between the `posPol` and `negPol` conditions for relative adjectives (`relAdj`) in the positive construction (`pos`) is plotted in the top leftmost plot of Figure 4. The comparison between the `posPol` and `negPol` conditions is similar to the hypothetical comparison between Michigan and Ohio students outlined above. On the x-axis we plot the abstract continuous $(-\infty,+\infty)$ evaluativity scale. The probability density curve corresponding to the `posPol` condition is plotted as a continuous line, while the density curve corresponding to the `negPol` condition is plotted as a dashed line. The means of the two distributions are 0 and -0.2, as listed in the plot. The two triangles on the x-axis indicate where these means are located: we see that the two triangles are basically on top of each other, indicating that the `posPol` and `negPol` conditions are not really distinct. This means that, in Experiment 1, subjects did not treat sentences like *John is short* and sentences like *John is tall* as different with respect to their evaluativity.

The four vertical dotted lines in the plot are the thresholds separating the areas under a density curve into five regions: (*i*) a -2 region, which is the area under the curve that is to the left of the first threshold, i.e., to the left of the leftmost dotted line; (*ii*) the -1 region, which is the area under the curve between the first threshold and the second threshold; (*iii*) the 0 region, which is the area under the curve between the second and the third threshold; (*iv*) the 1 region, which is the area under the curve between the third and the fourth threshold; and finally (*v*) the 2 region, which is the area under the curve to the right of the fourth (rightmost) threshold.

Figure 4 provides plots for both adjective classes – `relAdj` in the first column, `absAdj` in the second column – and for all five construction types – `pos` in the first row, `objComp` in the second row, `objEq` in the third row, `subjComp` in the fourth row, and finally `subjEq` in the fifth row. Note that the thresholds (the dotted vertical lines) are in the exact same place in all the plots: their placement is estimated based on the whole data and their position is the one that fits/accounts for the data the best.

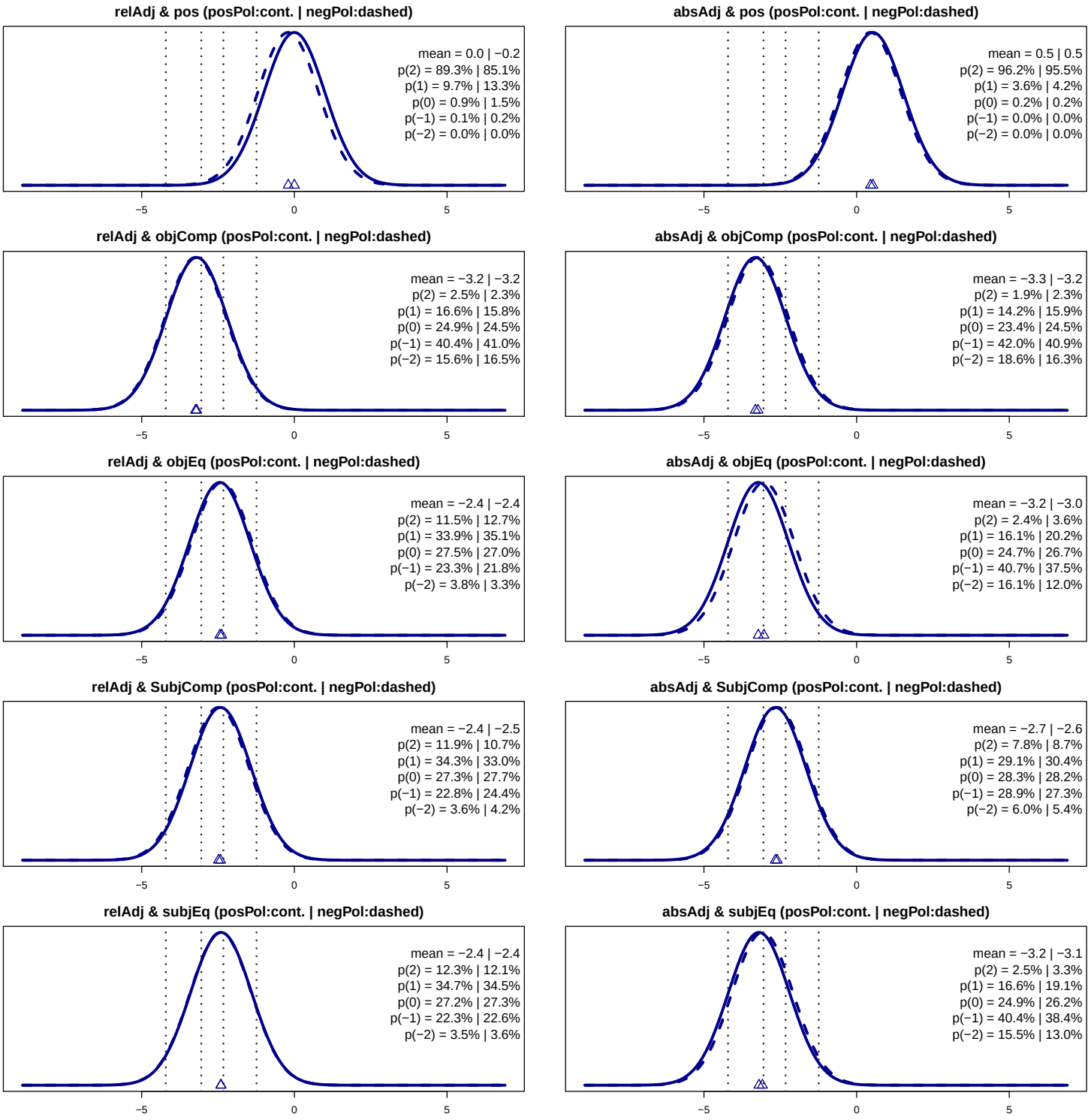

Figure 4: Experiments 1 and 2 – Plot of MLEs of the best mixed-effects ordinal probit model. The x axis is the abstract continuous $(-\infty,+\infty)$ evaluativity scale (only the $(-10,10)$ range is plotted). The probability density curve corresponding to `posPol` is plotted as a continuous line; the density curve corresponding to `negPol` is plotted as a dashed line. The four vertical dotted lines are thresholds separating the areas under a density curve into five regions: (*i*) the -2 region: the area under the curve that is to the left of the leftmost dotted line; (*ii*) the -1 region: the area between the first and second threshold; (*iii*) the 0 region: the area between the second and third threshold; (*iv*) the 1 region: the area between the third and fourth threshold; (*v*) the 2 region: the area to the right of the fourth (rightmost) threshold. These areas give the probabilities of the five possible outcomes -2, -1, 0, 1 and 2; they are listed in each plot for both `posPol` (on the left) and `negPol` (on the right).

Once estimated, the thresholds are fixed, so probabilities, i.e., the areas of the five regions under a curve, can vary only if the mean/center of the curve (represented as a triangle in Figure 4) moves to the left or to the right. The further to the right a density curve is, the more evaluative the corresponding condition. This is because moving a curve to the right increases the area of the rightmost regions, i.e., the probability of getting a 2 or a 1 answer. The further to the left a density curve is, the less evaluative the corresponding condition. This is because moving a curve to the left increases the area of the leftmost regions, i.e., the probability of getting a -2 or a -1.

As listed in the top leftmost plot in Figure 4, the probability of getting an answer of 2 for a positive construction with a positive antonym in the relative adjective condition is very high, namely 89.3%; this is because region 2 (the region to the right of the rightmost threshold) covers most of the area under the continuous (`posPol`) curve. The dashed (`negPol`) curve is slightly displaced to the left because its mean is -0.2, so its corresponding region 2 covers slightly less of the full area – only 85.1% percent. Thus, the probability of getting an answer of 2 for the `negPol` condition is 85.1%, which is slightly less than the corresponding probability of 89.3% for the `posPol` condition.

Technically speaking, the model whose maximum likelihood estimates (MLEs) are plotted in Figure 4 is a mixed-effects ordinal probit regression model, with fixed effects for adjective class, construction type, adjective polarity and all their 2-way interactions (but not the 3-way interaction between adjective polarity, adjective class and construction type), and intercept random effects for subjects and items. Likelihood ratio tests (LRTs) determined this to be the best model, i.e., the model that strikes the best balance between complexity (number of estimated parameters) and data fit. As we already indicated, the MLEs, their standard errors and corresponding p-values (if less than 0.05) are provided in Appendix B.4; these estimates provide support for the generalizations we listed in (27). The same appendix also provides more details about the LRTs we conducted. Finally, to intuitively see how close the model fits the data, we encourage the reader to compare the estimated (model-based) probabilities provided in Figure 4 and the corresponding empirical probabilities listed in Appendix B.3.

Analyzing evaluativity in terms of ordinal probit models assumes that the ordered categorical responses provide an imperfect image of an underlying continuous evaluativity dimension in the same way that grades provide an imperfect image of an underlying continuous scholarly ability. If this seems undesirable, the reader is encouraged to think of the underlying continuous probability distribution as a mathematical tool that enables us to statistically analyze our data while respecting its ordered categorical nature – in much the same way that we can think of possible-world semantics as a tool to analyze intensional phenomena in natural language without necessarily attributing psychological reality to possible worlds.

3.2 Discussion

The generalizations presented in §3.1.2 (see (27)) are supported by the ordinal probit model above. In what follows, we discuss these results and their theoretical consequences. We begin in §3.2.1 by discussing the (perhaps surprising) lack of a polarity effect. In §3.2.2 we discuss the unsurprising strong evaluativity of the positive construction, and in §3.2.3 we discuss the relative/absolute distinction that was the focus of this study. We end by discussing the comparison construction effects, which form the basis for Experiment 3.

3.2.1 The lack of a polarity effect

Recall that relative antonyms are reported to differ in their evaluativity in a complicated way (see Table 1 in §2.2.2). An adjective and its antonym are reported to pattern together with respect to evaluativity in two out of three tested degree constructions: the positive construction (where both antonyms are evaluative) and the comparative construction (where neither antonym is evaluative). An antonym and its adjective are expected to contrast in evaluativity only in the equative construction.

The first generalization reported above for Experiments 1 and 2 is that there is no main effect of adjectival polarity, for either the relative or absolute adjective pairs. We see a borderline significant main effect of polarity for relative adjectives in the positive construction (i.e. in the Experiment 1 data), but it goes in the unexpected direction: negative polarity antonyms seem to be less evaluative than their positive polarity counterparts. But because this effect is only borderline significant (the p-value is close to but slightly larger than 0.05; see appendix B.4), and no such effect was observed in the pilot experiment (see appendix A), we will conservatively take this effect to be noise and will not discuss it further.

The lack of a polarity effect is consistent with the across-the-board intuitions reported in Table 1 (with equative constructions being the sole exception). Numerically, we see that this contrast is reflected in our Experiment 2 data (i.e., for absolute adjectives) – see the slight displacement to the right of the dashed density curves in the `absAdj & objEq` and `absAdj & subjEq` plots in Figure 4. But this provides almost no evidence for or against the contrast reported in the theoretical literature. Follow-up experiments focused on polarity contrasts in equative and (possibly separately) positive constructions are needed before we can decide if our experimental task was not discriminating enough to detect the effect or if there is actually no effect.

3.2.2 The strength of evaluativity across constructions

For both relative and absolute adjectives, the positive construction tested as significantly more evaluative than the other constructions. One way to quantify this is to look at how probable it is to receive a rating of 2 or 1. For the positive construction, if we sum over $p(2)$ and $p(1)$, we get very close to 100% probability, i.e., the positive construction is basically certain to receive a rating of 2 or 1.[15] For all the other constructions, this probability is much lower: it never exceeds 20% for `objComp` and it is at most in the high 40%'s for the other 3 constructions. The overwhelming strength of the effect for the positive construction makes it more difficult to compare the potential evaluativity of the more complex comparison constructions.

There are a few plausible explanations for the strength of this effect. It is possible, as is traditionally assumed, that only the positive construction is evaluative (Bartsch and Vennemann 1972, Cresswell 1976, Kennedy 1999), and the perceived evaluativity of any other construction is insignificant or mistaken. Our statistical analysis does not present strong evidence against this claim but suggests that other adjectival constructions are somewhat evaluative and that subjects are sensitive to differences in evaluativity exhibited by these other constructions.

But if we take seriously the hypothesis that non-positive constructions are to some extent evaluative, how can we explain the large asymmetry in evaluativity between these constructions and the positive one? In particular, in view of the previous linguistic literature, we expected equative constructions (at least with negative polarity adjectives) to be basically as evaluative as the positive construction.

One possible explanation for the much stronger evaluativity of the positive construction is that its evaluativity is at-issue / truth-conditional while the evaluativity for the other constructions is presupposed / not at-issue. It's possible that the Bierwisch Test primarily targets at-issue evaluativity, and cooperative experimental subjects – who are sensitive to the distinction between asserted and presupposed content – respond accordingly.

An alternative explanation is that the strength of evaluativity for positive constructions is the result of the particular methodology employed here: positive constructions trivially pass the Bierwisch Test (the Bierwisch Test amounts to a test of whether $p \Rightarrow p$), while other constructions can only pass the test non-trivially. Thus, it is possible that the strength of the effect of evaluativity in the positive construction is an artifact of this methodological asymmetry.

But we see a significant difference between relative and absolute adjectives in the positive construction: absolute positive constructions are more evaluative than relative positive constructions – see Figure

[15] The breakdown for the four conditions is as follows: for `relAdj & posPol`, $p(2)$ and $p(1)$ total 99%; for `relAdj & negPol`, they total 98.4%; for `absAdj & posPol`, they total 99.8%; and for `absAdj & negPol` they total 99.7%.

4 above, where the absAdj means for the pos construction are about 0.5, while the means for relAdj & pos are 0 or thereabouts. And this makes the methodological asymmetry explanation (on its own) unlikely: the fact that there is a significant difference between absAdj and relAdj indicates that subjects were not treating the Bierwisch Test as a trivial inference in the case of the positive construction; instead, the contrast between absolute and relative adjectives reflects independently observed differences between the types of standards employed by the two types of adjectives (Rotstein and Winter 2004, Kennedy and McNally 2005).

3.2.3 The relative/absolute distinction

The Kennedy/McNally theory is that relative and absolute adjectives differ with respect to the structure of the scale they're associated with. In particular, that the scales associated with relative adjectives (tallness, heaviness) are unbounded, while the scales associated with absolute adjectives (opacity, transparency) are bounded, intrinsically, at one or more ends of the scale.

In part based on these observations, and in part based on an investigation of the interaction between these adjectival classes and vagueness, Kennedy (2007) argues that relative adjectives invoke contextually valued standards, while the standards invoked by absolute adjectives are their lexicalized bounds. This is his Principle of Interpretive Economy (19). Because relative adjectives have no conventional meaning that could correspond to a standard, the standards invoked by relative adjectives (but not absolute adjectives) vary from context to context. If this is right, then an account of the experimental results that emphasizes the role of context (or a lack of context) predicts that relative and absolute adjectives will behave differently. And this is in fact what we see.

In the positive construction condition, in which subjects were asked whether $p \Rightarrow p$, the entailment pattern between positive constructions with absolute adjectives was judged more acceptable than the entailment pattern between positive constructions with relative adjectives ($\beta = 0.53, SE = 0.20, p = 0.008$ – see the main effect for absAdj in the model estimates reported in Appendix B.4). That is, subjects were more likely to assent to (28-a) than (28-b).

(28) a. Bert's book is complete. ⇒ Bert's book is complete.
b. Linda's car is fast. ⇒ Linda's car is fast.

The raw data is as follows: for relAdj&posPol, 84% of the ratings were 2 and 13% of the ratings were 1, while for absAdj&posPol, 93% of the ratings were 2 and 5% of the ratings were 1 (for more information, including the actual counts, see the table in Appendix B.3). This is a remarkable result if subjects are treating the positive construction only as an instance of trivial entailment.[16] The putative difference between relative and absolute adjectives in terms of their context sensitivity can help explain the difference, though: because the standards invoked by absolute adjectives are just their bounds, absolute positive constructions are not context-sensitive in the way that relative ones are. That is, if it's the case that Bert's book is complete in one context, it's complete in all contexts. But this is not necessarily the case if Helen's book is good.

This strongly suggests that subjects are aware of contextual sensitivity in these constructions, and that they are less likely to accept the Bierwisch Test in the case of a contextually sensitive positive construction than they do in the case of a contextually insensitive construction. Importantly, it also suggests that subjects are considering contextual variability when they are asked about the reliability of a sentence's entailment, even when their judgments are elicited with respect to a specific context.

[16] A reviewer wonders if these results are instead some methodological artifact. It is true that binary category membership (like for the predicate *is a prime number*) is manifested as gradable in experiments that use a Likert scale (Armstrong et al. 1983, Hansen and Chemla 2013). But the reported differing degrees to which e.g. a number is considered prime is generally viewed as representative of our intuitions about category membership. We see no reason to think otherwise here. Our goal is to explain why and how the context-sensitivity of this methodology resulted in subjects differentiating in this way between relative and absolute adjectives.

Other constructions also show evidence of a difference between relative and absolute adjectives. While subjects did not treat relative and absolute adjectives differently in comparatives, they did treat them differently in equatives: equatives with relative adjectives were judged to be more evaluative than equatives with absolute adjectives. This means that the entailment pattern exemplified in (29-a) was judged as significantly more acceptable than the entailment pattern exemplified in (29-b), for both subjects and objects of equatives.

(29) a. Bill's cactus is as tall as Amy's cactus. ⇒ Bill's/Amy's cactus is tall.
b. Joe's recitation of the poem is as perfect as Danielle's recitation of the poem.
⇒ Joe's/Danielle's recitation of the poem is perfect.

Recall from §2.1 a reported asymmetry between the subjects and objects of equatives: projection tests determined that the evaluativity presupposed by equative constructions seemed to be associated with the objects, but not the subjects, of equatives. Despite this asymmetry, the truth conditions of an equative equate one height with another, which effectively transfers the evaluativity of one individual to the other individual. This is confirmed by the near equality of the `subjEq` and `objEq` conditions.

In sum, subjects clearly differentiated between relative and absolute adjectives, but they did so in a complicated way. They treated absolute adjectives as significantly more evaluative in positive constructions, but relative adjectives as significantly more evaluative in equatives. And they didn't differentiate between relative and absolute adjectives in comparatives. It seems plausible that this result evaded prior notice because, unlike reports of speaker's intuitions about evaluativity, the experimental task we designed was both fine-grained and precise enough to determine if two evaluative constructions differ in their degree of evaluativity (or the degree to which speakers feel comfortable characterizing evaluativity as a reliable entailment of a construction across contexts).

It is not clear how to reconcile these two differences between relative and absolute adjectives. One prominent difference between positive constructions and equatives is that the former involves one individual and its measure, while the latter involves two individuals and their measures. The Kennedy/McNally theory predicts a positive construction formed with an absolute adjective to be evaluative because the standard of comparison invoked is the adjective's maximum. But it's possible that the introduction of two relevant individuals makes the interpretation of an absolute-adjective equative more vulnerable to contextual considerations, thereby weakening the relationship between the absolute adjective's maximum and evaluativity in the context of evaluation.

In particular, it might be that in the case of absolute adjectives, there is a (pragmatic) competition between two different equative strategies, the coordination strategy in (30-a) (Haspelmath and Buchholz 1998) and the familiar degree-quantifier strategy in (30-b).

(30) a. The glass and the vase are empty.
b. The glass is as empty as the vase.

While (30-a) equates individuals with respect to an absolute property, the degree-quantifier strategy in (30-b) equates degrees of emptiness. In the case of absolute properties (and thereby maximal standards), the choice between these two constructions might bear on the evaluativity of the sentence: an utterance of (30-b) might be taken to implicate that the maximum standard is not met in the context of interpretation.

In conclusion, it's clear that subjects calculate evaluativity differently for relative and absolute adjectives. In the case of positive constructions, which are overwhelmingly judged to be evaluative, they seem to do so in a way that is predicted by the Kennedy/McNally approach: positive constructions formed with absolute adjectives are even more likely than those formed with relative adjectives to be judged as evaluative. This is presumably because the evaluativity of these constructions is conventionalized or lexicalized in a way that it is not for relative adjectives. However, the (less evaluative) equative constructions displayed the opposite trend. While we have only speculated about this potential dif-

ference between positive constructions and equatives, these curiosities about the role of context in this experiment and in the calculation of evaluativity generally – as well as the result about comparative constructions discussed immediately below in §3.2.4 – forms the basis for Experiment 3, discussed in §4.

3.2.4 Evaluativity in comparison constructions

This section discusses one final major result of Experiments 1 and 2. Despite the dwarfing effect of the positive construction, there is still reason to think that subjects treated the two comparison constructions – comparatives and equatives – as evaluative, at least to some extent.

Across adjective types, participants judged the subject position of a comparative to be more evaluative than the object position (or the subject/object positions in equatives). Specifically, the subject position of comparatives entails the corresponding positive construction to a significantly larger extent than the object position of comparatives: $p(2) + p(1)$ is around 40% for `subjComp` and less than 20% for `objComp`. This means that subjects were significantly more likely to accept the entailment in (31-a) than the one in (31-b) (or the entailments for either argument of the equative).

(31) John is taller than Bill.
  a. $\Rightarrow$ John is tall.
  b. $\Rightarrow$ Bill is tall.

This result is in direct conflict with intuitions reported in the theoretical literature, illustrated in (32) (and discussed in §2.1).

(32) a. John is shorter than Bill. $\not\Rightarrow$ John/Bill is short.
  b. John is shorter than Bill, but they're both tall.

But, interestingly, this result is consistent with some preliminary experimental investigations of evaluativity in the psychology and experimental linguistics literature. Psychologists interested in deductive reasoning as a cognitive process in children and adults have tested the processing and interpretation of comparatives (Piaget 1928, Donaldson 1963, Clark 1969a,b, Flores D'Arcais 1970, Clark 1972, Higgins 1977). For example, although Clark's focus was on the speed at which the subjects correctly deduced transitive relationships encoded in multiple comparative constructions, he also tested for evaluativity (which he referred to as 'contrastiveness'), focusing almost exclusively on *better* and *worse* comparatives.[17]

Clark observed that for a comparative of the form 'A is P-er than B,' subjects are likely to conclude that A is P, but not that B is P (p397). In other words, subjects are more likely to interpret a comparative as characterizing the external argument as evaluative. This claim is strengthened by work in Flores D'Arcais (1970) (who argues that this 'positional asymmetry' is restricted to comparatives formed with positive antonyms). The result seems especially clear in the testing of children. Clark reports an earlier study conducted by Piaget, in which children were presented with the following problem:

(33) Edith is fairer than Suzanne; Edith is darker than Lili. Which of the three has the darkest hair?

Piaget's description of their responses is as follows:

> "It is as though [the child] reasoned as follows: Edith is fairer than Suzanne so they are both fair; and Edith is darker than Lili so they are both dark. Therefore Lili is dark, Suzanne is fair, and Edith is between the two. In other words, owing to the interplay of the relations included in the test, the child, by substituting the judgment of membership (Edith and Suzanne are

[17] We refer to these results as 'preliminary' because they tend not to control for known complications of evaluativity in the current linguistics literature, namely the difference between analytic and synthetic morphology, relative and absolute adjectives, etc. See Rett (2008b, 2015) for an overview of these issues.

> "fair," etc.) for the judgment of the relation (Edith is "fairer than" Suzanne), comes to a conclusion which is exactly opposite of ours." (Piaget 1928: 87)

Even more striking, the evaluativity in comparatives is so compelling that when the children were presented with a series of comparatives that differed in polarity, they were forced to contradictory conclusions.

> The children in Donaldson's (1963) studies often made other errors as a result of their comprehension of propositions as base strings. For example, children were given the following [...] problem: "Dick is shorter than Tom. Dick is taller than John. Which of these three boys is tallest?" Even though the problem explicitly states that there are three boys, many children assumed there were four. They said there were two Dicks – a tall one and a short one – following the analysis of the base strings. One girl's solution to the above problem was, "This Dick [second premise] is tallest, John is next tallest, Tom is third and then it's Dick [first premise]". (Clark 1969b: 399)

This sort of error is apparently very frequent: it accounts for about 70% of errors in Donaldson's study.

In contrast, Higgins (1977) observed no difference in evaluativity between the subject and object position. (Like the others, he does not test for evaluativity in any other adjectival construction, although he did expand his stimuli to include relative and extreme adjectives, like *obese*.) But his method departs from the Bierwisch Test: instead of asking about the meaning of a sentence, he asks about its acceptability (whether it "sounds right") in contexts in which the evaluativity is implausible, e.g. for sentences like *A feather is heavier than a snowflake*. In every case, judgments of acceptability relied on the subjects' world knowledge in addition to their linguistic knowledge. It therefore seems reasonable to attribute the contrast between his results and the others' to the fact that others used some version of the Bierwisch Test, while Higgins tested for evaluativity by seeing which constructions were (in)compatible with contextually insensitive real world knowledge of evaluative properties.

Getting back to the present study: the results of Experiments 1 and 2 suggest that participants consider the subjects of comparatives to be evaluative, in contrast to the objects of comparatives and even to (either argument of) equative constructions. These results directly contradict intuitions reported in the linguistics literature and run up against theories designed to predict those intuitions.[18] They do, however, seem to replicate some limited previous experiments in the psychology literature – in particular those in Piaget (1928), Clark (1969a,b, 1972) and Flores D'Arcais (1970) – but arguably only those running some experimental variant of the Bierwisch Test.

What accounts for this surprising result? On one hand, it's possible that the reported intuitions – although widespread – are mistaken. On the other hand, it's possible that the Bierwisch Test, at least in the context of the comparative construction, does not appropriately target evaluativity. In what follows, we present a hypothesis along the lines of this second possibility; in particular, that participants in Experiments 1 and 2 judged the subjects of comparatives to be evaluative because they interpreted comparatives – but not equatives – as creating a context in which the most salient contextual standard is the measure of the individual denoted by the object. Subjects necessarily exceed this measure in comparatives (which encode the $>$ relation), but not in equatives (which encode the $\geq$ relation). Accordingly, the subject positions of comparatives are thus judged to be evaluative, but with respect to the object argument, not some other contextual standard. Experiment 3 was designed to directly test this hypothesis, which we expand upon below.

Our hypothesis is in a way expected given the analysis of adjectival vagueness in Barker (2002) as a set of contextual restrictions on degree variables, which is incrementally updated in discourse. He

[18] For example, Barker (2002: 2) (who equates evaluativity with the ability to have a metalinguistic interpretation) says explicitly: "Measure phrases and comparatives have no metalinguistic side-effects: declaring that Bill is six feet tall, or that Bill is six inches taller than Feynman, reveals nothing about what counts as tall. On the analysis here, the impossibility of using measure phrases or comparatives to negotiate vague standards follows from the truth conditions of the constructions."

argues that in addition to having presuppositional and assertive content, these 'vague' constructions carry a third type of entailment – sharpening – which he defines as "what happens to the status of vague predicates as discourse constrains possible ways of resolving that vagueness" (p. 4; see also Kamp 1975, Klein 1980, Pinkal 1989). The sharpening components of a construction can also be thought of as metalinguistic meaning.

For example: in a context in which it's clear what the relevant standards of tallness are, the positive construction *John is tall* carries descriptive meaning; it counts as contributing information about John's height. In contrast, Barker describes a context in which Feynman's height is clear, but an interlocutor is interested in knowing what counts as tall in the U.S. In such a context, the positive construction – by virtue of its being evaluative – has a metalinguistic contribution; it carries information about the relevant standard of tallness, rather than Feynman's height. As Barker says:

> "In fact, assuming that *tall* means roughly 'having a maximal degree of height greater than a certain contextually-supplied standard', I haven't even provided you with any new information about the truth conditions of the word *tall*. All I have done is given you guidance concerning what the prevailing relevant standard for tallness happens to be in our community; in particular, that standard must be no greater than Feynman's maximal degree of height" (Barker 2002: 2).

As Barker presents it, an evaluative construction (or, more generally, a context-sensitive construction) is descriptive in a context in which the contextual standard is known, but meta-linguistic or sharpening in a context in which the contextual standard is unknown.[19]

This dichotomy seems relevant for reconciling the intuitions reported in the theoretical literature – that the comparative is not evaluative, while the equative is – with the experimental results presented here and in the prior psychology literature that the subject of the comparative is evaluative.

In the theoretical literature, evaluativity judgments are often reported in a particular context; in §2.1, we presented them in a context in which John is either extremely tall or extremely short. In experimental studies, on the other hand, evaluativity judgments are extracted in the absence of any contextual information that could inform the speaker about the relevant standard of e.g. tallness. In the absence of any other information about the context of evaluation, the subject can only infer from the comparative in (31) that (*i*) the context of evaluation contains two individuals, John and Bill; (*ii*) those individuals are strictly ordered with respect to their height; and (*iii*) John's height exceeds Bill's height. The Bierwisch Test relies on the participants' ability to reason about the meaning of sentences across contexts, but it's possible that subjects instead treated this minimal context as the only relevant context of evaluation. And it seems plausible that in a world in which there are only two individuals, and those individuals are strictly ordered with respect to height, one individual will count as tall, and the other one as not tall (see van Rooij 2011).

In contrast, equatives do not strictly order individuals with respect to a dimension of measurement. Therefore, if the above 'sharpening' hypothesis is correct – that is, if subjects are treating the object of comparatives as the relevant standard in the context of interpretation – we also expect that equatives should *not* exhibit a positional asymmetry with respect to evaluativity since equatives do not strictly order their two individuals. And this is in fact what we see: within each of the two `relAdj` and `absAdj` class, `subjEq` and `objEq` constructions are very similar.

It is then possible that a comparative uttered in a context with a clear standard is not evaluative (its descriptive use), while a comparative uttered in a context with no clear "prevailing standard" is (its metalinguistic use). In the latter context, implicit comparatives like the degree comparatives studied here would function more like explicit comparatives ((34-a), discussed in Kennedy 2009, Sawada 2009, Kubota and Matsui 2010) or conjoined comparatives ((34-b), discussed in Stassen 1985).

[19] Although Barker (2002: 10) cautions that this is an idealized perspective: "in the normal situation, an utterance of a sentence like [*Feynman is tall*] simultaneously operates at both a descriptive and a metalinguistic level."

(34) a. Relative/Compared to Bill, John is tall.
b. John is tall and Bill is short/not tall.

In fact, as a reviewer points out, there is independent experimental evidence that subjects are willing to partition ad-hoc comparison classes for the purpose of interpreting evaluative constructions in the absence of explicit information about the standard. Syrett et al. (2010) demonstrates this for children's and adult's interpretation of positive constructions; Tribushinina (2011), Tribushinina and Gillis (2012) demonstrate the same for children's use of degree modifiers. The present study can be taken as evidence that this process is at work outside of evaluative constructions: a variety of degree constructions are interpreted as informing subjects about the nature of the comparison class and, in so doing, the nature of the relevant standard of comparison.

It is important to note that Barker (2002) expressly assumes that degree comparatives are not evaluative, even in the sense we're suggesting, and Kennedy (2009), Sawada (2009) and Kubota and Matsui (2010) all consider evaluativity to be a characteristic difference between explicit (degree) comparatives (which allegedly lack it) and implicit comparatives like (34-a) (which are evaluative for the subject, *John* in this case).

Experiment 3, to which we turn in the next section, is designed to test the extent to which the 'sharpening' hypothesis is a plausible explanation of the discrepancy between the traditional description of evaluativity in the theoretical literature and our experimental results (or the experimental results in the psycholinguistic literature). If Experiment 3 supports our sharpening hypothesis, it seems that the traditional description of evaluativity is not exactly right / is lacking in nuance. In particular, evaluativity might be detectable in contexts of evaluation with no clearly defined standard (like comparatives) if our measurement procedure is sensitive enough, i.e. not exclusively based on introspective judgments.

3.3 Interim summary

Experiments 1 and 2 comprise a comprehensive study of evaluativity across representative degree constructions and adjective classes. The experimental method employed here is an adaptation of the Bierwisch Test: a Detective utters a degree construction (e.g. *Martha is as strong as Bertha*) and asks participants to rank on a 5-point Likert scale how justified his interlocutor is in concluding that a corresponding positive construction (e.g. *Bertha is strong*) is true.

While this test amounted to something that looked like a trivial entailment in the case of the positive construction – i.e., a judgment about whether p entails p – our results in this condition varied significantly across adjective class, suggesting that participants may have relied heavily on contextual cues to differentiate between different uses of the same construction within a single context. This is one of several ways in which Experiments 1 and 2 revealed significant considerations of context in participants' judgments of the entailments of an utterance.

In addition to illuminating the role of context in judgments of evaluativity, the experimental design added a level of nuance to the study of evaluativity not present in theoretical reports of speaker intuition: its introduction of a Likert scale into the Bierwisch Test allowed for participants to report graded notions of evaluativity. The result is a study of evaluativity that is not categorical (although this doesn't entail that we need a notion of evaluativity that isn't categorical; Hansen and Chemla 2013). These gradable responses are particularly useful in comparing the behavior of different adjectival classes within a particular construction – like the positive construction – which has been claimed to be evaluative regardless of adjective type. It also has the potential to better reflect different types or sources of evaluativity. In particular, we were additionally interested in examining the distribution of evaluativity across degree constructions (in particular, comparatives and equatives) and across adjective classes (in particular, relative and absolute adjectives).

With respect to the relative/absolute distinction, the experiment produced two interesting results: first, that participants judged absolute-adjective positive constructions to be significantly more eval-

uative than relative-adjective positive constructions; and second, that they judged relative-adjective equatives to be significantly more evaluative than absolute-adjective equatives. While the proposal in Kennedy and McNally (2005) predicts these classes of adjectives to behave differently with respect to evaluativity, we know of no theory which predicts a construction-specific difference.

Perhaps the most striking result of Experiments 1 and 2 is the relative evaluativity of the subject position of comparative constructions. In the linguistic literature, both positions of the equative are reported to be evaluative, in contrast to the comparative where neither position is claimed to be evaluative. But in our experiments, the subject of the comparative was judged to be more evaluative than either argument of the equative, which in turn were judged to be more evaluative than the object of the comparative.

While this result does not seem to be compatible with the intuitions reported in the theoretical literature, it is consistent with some early experiments in the psycholinguistic literature (in particular, those who employ some version of the Bierwisch Test). In §3.2.4, we hypothesized that this unexpected result is the consequence of an unforeseen interference of the null context in which the stimuli was presented. In particular, that participants used the comparative (but not the equative) to determine the relevant standard of evaluation. A context consisting only of two strictly ordered individuals is quite naturally divided into a positive and negative extension (the subject and object, respectively), while a context consisting of only two non-strictly ordered individuals is not so naturally divided.

If this is right, it seems as though we have reason to differentiate between two types of evaluativity that can be associated with a given construction: evaluativity that is intrinsic to the construction (part of its descriptive content, either at-issue or not-at-issue) and evaluativity that arises in particular contexts as the result of the construction's descriptive content plus contextual considerations. The former is the traditional notion of evaluativity, while the latter looks somewhat like what Barker (2002) referred to as the metalinguistic use of evaluative constructions. The evaluativity conferred on the subject of a comparative is like Barker's metalinguistic use of positive constructions in that it serves to value the standard of comparison. It is, however, unlike metalinguistic uses of positive constructions in that comparatives, in contrast to positive constructions, are generally not considered to be evaluative in other contexts.

This is therefore an interesting result in a number of respects: if our explanation is correct, the apparent evaluativity of the subject position of comparatives has the potential to shed light on the how context-sensitive expressions are valued and restricted (in Barker's words, sharpened) from context to context. In addition to informing the study of evaluativity, a better understanding of the interpretation of context-sensitive expressions could be extended to other context-sensitive phenomena as well as models of discourse and context update.

4 Evaluativity and context-sensitivity

In this section, we report the results of an experiment explicitly designed to test the hypothesis presented above: namely, that the reported evaluativity of the subject of comparatives is the result of the effect the comparative has on shaping / sharpening the relevant context of evaluation (rather than the evaluativity intrinsic to this construction). Experiment 3 is a version of Experiments 1 and 2 in which three rather than two individuals are introduced, and there is a strict ordering between these three individuals. We conclude in §5 by discussing the broader implications of this entire series of experiments and directions for future work.

4.1 Method

The main goal of Experiment 3 was to test whether the evaluativity of comparative constructions was sensitive to the context implicitly set up by the comparative in a way that equative constructions are not. In particular, we wanted to test two competing hypotheses about the evaluativity associated with the subject of the comparative in Experiments 1 and 2: *(i)* subjects of comparatives were considered

evaluative because the subject position of comparatives is intrinsically evaluative, like the positive construction; or *(ii)* subjects of comparatives were considered evaluative because of a contextual effect of the comparison class implicitly set up by the comparative: namely, two entities strictly ordered with respect to the scale contributed by the adjective, the higher of which must *per force* count as exceeding the contextual standard. To test these competing hypotheses, we added a second comparison construction to the stimuli, as in (35).

(35) The Detective reported to the Police Chief: "Maria is taller than Sophie and Sophie is taller than Joe."

a. First subject targeted (subjComp1): The Chief concluded from this that Maria is tall.
b. Second subject targeted (subjComp2): The Chief concluded from this that Sophie is tall.

While Experiments 1 and 2 tested evaluativity in the subject and object positions of a single comparison construction, Experiment 3 tested evaluativity in the subject positions of each of the two conjoined comparison constructions: (35-a) shows an example that tests the subject of the first comparative, while (35-b) shows an example that tests the subject of the second comparative (which in every case was also the object of the first comparative).

The first hypothesis (evaluativity is intrinsic to the subject position of comparatives) predicts that there should be no difference between the evaluativity of (35-a) and the evaluativity of (35-b) above, modulo the discourse effects (if any) associated with the fact that the subject in (35-a) linearly precedes the subject in (35-b). In contrast, the second hypothesis (evaluativity due to the implicitly constructed comparison class) predicts that only (35-a), but not (35-b), should exhibit clear evaluativity effects.

To factor out the presumed effects of linear precedence, and to have a general baseline against which to assess the evaluativity of subjects in comparatives, we also presented participants with equatives in the same kind of three-entity contexts:

(36) The Detective reported to the Police Chief: "Maria is as tall as Sophie and Sophie is as tall as Joe."

a. First subject targeted (subjEq1): The Chief concluded from this that Maria is tall.
b. Second subject targeted (subjEq2): The Chief concluded from this that Sophie is tall.

We expect to see no difference between (36-a) and (36-b) under either of our two hypotheses since the presence of a strict ordering of the kind induced by comparatives but not equatives is a necessary precondition for both hypotheses.

Since the effect of `subjComp` relative to `objComp` was most clearly visible for relative adjectives, Experiment 3 focuses only on this adjectival class. Experiment 3 stimuli were formed with basically the same relative adjectives and types of entities as the Experiment 1 stimuli.

Each of the 32 items was passed through 8 conditions: 2 adjectives polarities (`posPol`, `negPol`) $\times$ 4 construction types (`subjComp1` (35-a), `subjComp2` (35-b), `subjEq1` (36-a), `subjEq2` (36-b)). Every participant saw each of these items exactly once, either with the `posPol` or the `negPol` adjective, and in one of the 4 construction types, for a total of 32 stimuli (excluding fillers). The adjective polarity & construction type combination was randomly selected for every one of the 32 items and rotated for every participant (Latin square design). Except for the fact that the Detective's reports involve a conjunction of comparison constructions, the method for Experiment 3 was identical to the method for Experiments 1 and 2.

In addition to the 32 experimental stimuli, there were 64 fillers that involved similarly 'transitive' scenarios involving three entities. The fillers came in pairs: the first member of the pair was expected to yield a high rating (justified conclusion), and the second member was expected to yield a low rating (unjustified conclusion). An example is given below:

(37) a. *John gave some money to Margaret and Margaret gave some money to Mike.* $\Rightarrow$ *Margaret got some money from John.*

b. *John gave some money to Margaret and Margaret gave some money to Mike.* ⇒ *Mike got some money from John*.

Thus, every participant responded to 96 stimuli (32 experimental stimuli + 64 fillers), the order of which was randomized for every participant. 45 participants (undergraduate students from UCSC) completed the study online on a UCSC-hosted installation of the IBEX platform. Since Experiment 3 did not use the (uncontroversial) positive construction as baseline / control, we used the fillers to filter out 2 participants that did not properly complete the experiment. The final number of participants is 43, for a total of $43 \times 32 = 1376$ observations.

4.2 Results

Graphical summaries of the data collected in Experiment 3 are provided in Figure 5 below. There are 2 plots in that figure. The leftmost plot corresponds to the positive polarity (`posPol`) adjectives, while the rightmost plot corresponds to the negative polarity (`negPol`) adjectives. For completeness, the data summaries plotted in Figure 5 are provided in Appendix C.2.

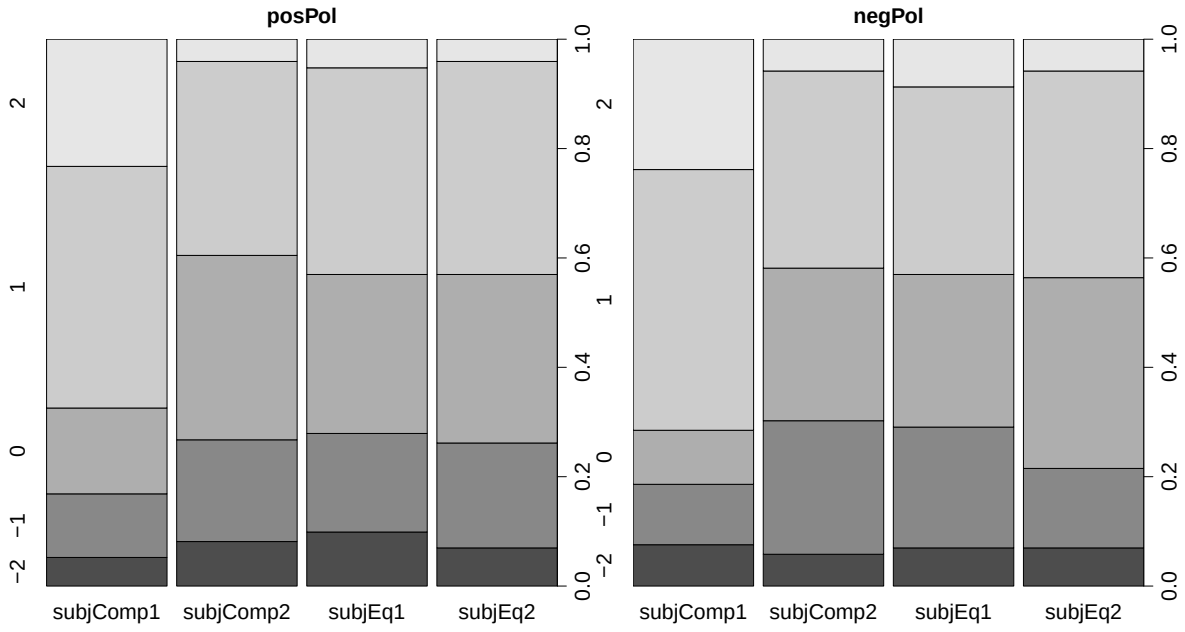

Figure 5: Experiment 3 – Graphical summaries

These plots suggest that the prediction made by our second hypothesis (evaluativity due to the implicitly constructed comparison class) is correct – see (38-a) below:

(38) a. The `subjComp1` condition is more evaluative than `subjComp2` condition (and also more evaluative than both equative constructions).
b. Once again, there seems to be no difference between positive and negative polarity adjectives across comparison constructions.

These generalizations are confirmed by the statistical analysis presented in the next section.

4.3 Statistical analysis and brief discussion

We estimate the same kind of models as before: mixed-effects ordinal probit regression models with intercept-only random effects for subjects and items. An LRT test shows that the full-fixed-effect structure model (main effects plus all two-way interactions) does not significantly reduce deviance relative

to the main-effects only model ($\chi^2 = 0.3, df = 3, p = 0.96$). We will therefore discuss the main-effects model from now on.

The estimates are listed in Appendix C.3 and plotted in Figure 6. We can clearly see that there really is no difference between `posPol` and `negPol` adjectives for any construction type, and that the means for `subjComp1` (around $0.9 - 1$) are higher than the all the other means (which are all around $0 - 0.2$). In particular, there is a statistically significant difference between `subjComp2` and `subjComp1` ($\beta = 0.93, SE = 0.089, p < 2 \times 10^{-16}$).

Thus, our hypothesis that subjects are more evaluative than objects in comparatives because participants implicitly construct a comparison class as they interpret comparatives uttered out-of-the-blue is supported by our follow-up Experiment 3.

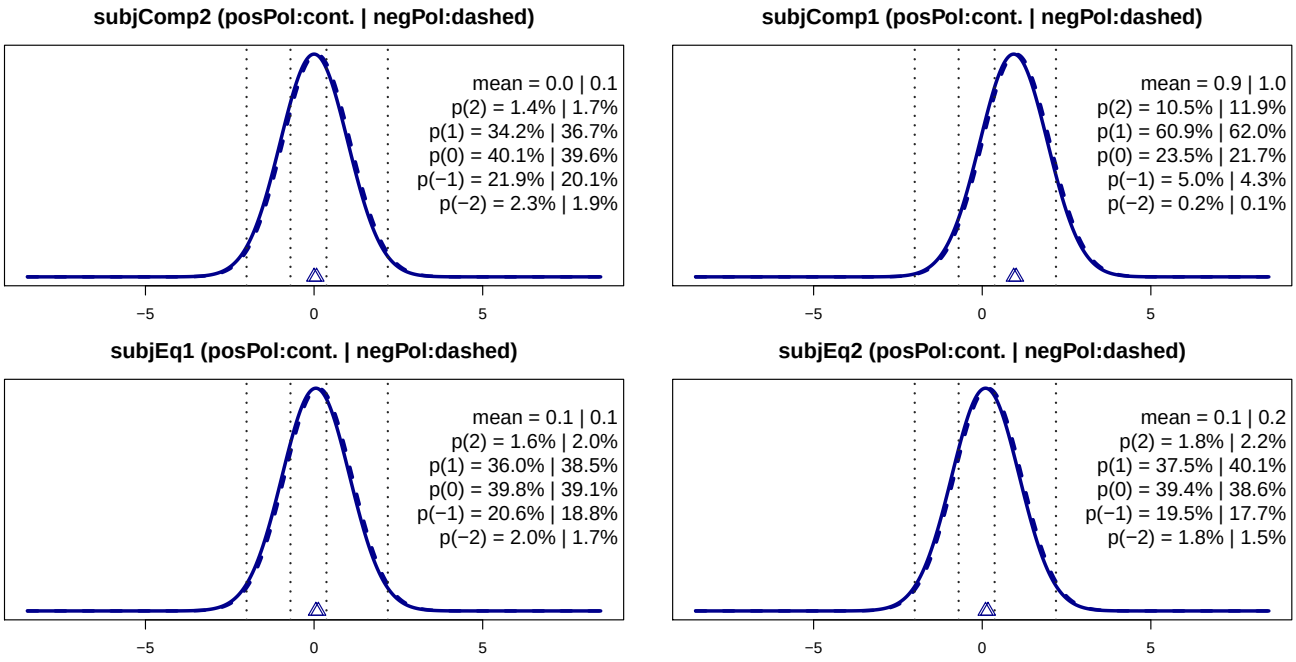

Figure 6: Experiment 3 – Plot of MLEs of the best mixed-effects ordinal probit model.

5 General discussion and conclusions

The goal of the present project was to quantify the distribution of evaluativity across adjectival constructions and classes. The results of Experiments 1 and 2 confirmed that positive constructions (e.g. *John is tall*) are overwhelmingly evaluative. They also confirmed that there is a fundamental difference between relative and absolute adjectives in terms of evaluativity, although the difference is not consistent across constructions: relative adjectives are significantly more evaluative in equative constructions, while absolute adjectives are significantly more evaluative in positive constructions. There is, as far as we know, no current theory of the relative/absolute distinction or of evaluativity that provides a clear and straightforward explanation of this difference.

Due to the fairly wide-ranging nature of the experimental investigation, we have been unable to draw more specific conclusions about, for instance, differences in evaluativity within the class of absolute adjectives, or the effect of adjectival polarity on the evaluativity of equative constructions. Since Experiments 1 and 2 have laid the groundwork for these more detailed investigations, we hope to follow up soon with more targeted studies addressing these issues.

The second half of this paper focused on the subject position of comparatives. Traditionally, neither argument of the comparative is characterized as evaluative: *John is taller than Mary* is intuitively true and

felicitous in contexts in which John or Mary fail to count as tall. In direct opposition to this reported intuition, the results of Experiments 1 and 2 showed that participants consistently endorsed an inference from comparatives (e.g. (39-a)) to the conclusion that the subject of the comparison qualifies as significantly *Adj* ((39-b)). They did not do so for the objects of comparatives, and they did so only to a lesser extent for the subjects and objects of equatives.

(39) a. John is taller than Mary.
b. John is tall.

We pointed out that this particular result is not entirely unfamiliar: several experimental studies in the psychology literature also report (in some form or another) that the subject of comparatives is evaluative (Clark 1969a,b, 1972, Flores D'Arcais 1970). In contrast to at least one other study (Higgins 1977), it seems to be the case that experiments reporting evaluativity for the subjects of comparatives (including our own Experiments 1 and 2) have in common a particular methodology: some variant of the Bierwisch Test.

To explain this result, particularly in light of our own intuitions and reports in the theoretical literature, we entertained two competing hypotheses. First, that the theoretical literature and the intuitions it reports have been mistaken: the subject position of comparatives really is associated with evaluativity, similar to the positive construction. And second, our 'sharpening' hypothesis, which contends that the evaluativity of subjects in comparatives is a consequence of the interaction between the descriptive content of the comparatives and the null context in which we presented our stimuli. In other words, the sharpening hypothesis contends that subjects assent to the inference from (39-a) to (39-b) because they are using the content of the premise – the truth conditions of the comparative – to help determine (or sharpen) the contextual standard relevant for the calculation of evaluativity.

The second (sharpening) hypothesis is more explanatory than the first: it explains the asymmetry between the two arguments of the comparative (recall that the object of a comparative is the least evaluative of all the constructions and positions we tested), and it also explains why participants are more willing to assent to the inference in (39) than they are for the subjects of equatives (e.g. *John is as tall as Mary... John is tall*). Equatives encode a non-strict ordering between individuals, so participants cannot use the truth conditions of equatives as a potential source of information about who or what counts as tall in the context of interpretation.

To test these two hypotheses, we ran a third experiment which differed from the first two in that its premise was a conjunction of two comparatives (and two equatives), as in (35) and (36). Instead of testing the subject and object of a single comparative, we tested the evaluativity of the subjects of the first and second comparison construction. If the subject of comparatives were intrinsically evaluative (the first hypothesis), we would expect these two subjects to be equally evaluative. If, instead, the results from Experiments 1 and 2 reflected participants' reasoning about the significance of the comparative for the purposes of calculating evaluativity, we would expect the subject of the first comparative to be more evaluative than the subject of the second comparative. This is because the former is clearly tall in a context of three individuals, while the latter is potentially in the 'extension gap' between the extension of *tall* and *not tall* in that context.

And this is, in fact, what we found. The subjComp1 condition was significantly more evaluative than the subjComp2, subjEq1 and subjEq2 conditions. Specifically, the evaluativity effect associated with the subjects of comparatives in Experiments 1 and 2 was neutralized when the context of interpretation included three strictly ordered individuals instead of two, at least for the individual that fell in the middle of the ordering. We thereby conclude that the 'evaluativity' associated with the subject position of comparatives in Experiments 1 and 2 reflected what participants were inferring about the context of evaluation, rather than that the subject position of comparatives is intrinsically evaluative (like it seems to be for the positive construction).

We will close by discussing three interesting consequences of this result. As discussed in §3.2.4,

the observation that evaluative constructions can be used to inform about the nature of the contextual standard in addition to about how individuals measure up to the contextual standard has been made in the previous literature. Barker (2002) referred to this use of evaluative constructions as 'metalinguistic', as opposed to their standard, descriptive use. At first glance, it seems as though the work in Barker 2002 gives us a good way of describing participants' treatment of subjects of comparatives in Experiments 1 and 2.

However, Barker's observations are subtly and interestingly different from the results highlighted here. His theory characterizes an adjectival construction (he focused on the positive construction) as receiving either a descriptive or metalinguistic interpretation in a particular context. If we're right, it seems as though the descriptive interpretation of an expression can in principle be used to affect the contextual content of that expression, contrary to Barker's predictions. In effect, this suggests that a sentence can simultaneously receive a descriptive and a metalinguistic interpretation. We hope to explore this further; a better understanding of the interaction of a sentence's compositional and contextual meaning will clearly serve to inform semantic theories of context-sensitivity.

A second interesting consequence of this study is the possibility that evaluativity is a gradable property rather than a categorical one. This is not a necessary interpretation of our results (see Hansen and Chemla 2013 for discussion of this issue with respect to other categorical semantic notions)[20] but pending a careful investigation of this vs. alternative hypotheses, we want to outline some of its consequences. The positive construction was overwhelmingly more likely to pass the Bierwisch Test than any other construction or position; but even within the positive construction, there was a significant difference between those formed with absolute adjectives (more evaluative) and relative adjectives (less evaluative). This runs contrary to standard intuitions reported in the theoretical literature, in which adjectival constructions are classified as either evaluative or non-evaluative (Bierwisch 1989, Rett 2008b, 2015).

But the gradient nature of evaluativity also seems to be incompatible with many accounts of context-sensitive phenomena generally. Broadly speaking, accounts of context-sensitivity in linguistic semantics and philosophy of language fall into two categories: semantic accounts in which context-sensitive constructions contain an 'unarticulated constituent' that corresponds, in the construction's denotation, to a free variable that is valued by context (Stanley and Szabo 2002, among others); and pragmatic accounts, which argue that contextual information is instead filled in by rational interlocutors trying to comprehend the speaker's meaning (Sperber and Wilson 1986, Cappelen and Lepore 2005, among others).

Semantic accounts of context-sensitivity have received particular attention for at least two reasons. First, they are closely related to 'contextualist' accounts of vagueness (Kamp 1981), which have been proposed to account for problems related to vagueness (e.g. the Sorites Paradox). Second, unlike pragmatic accounts, they predict that words/constructions are sharply divided into two categories: those that are context-sensitive (and whose denotations include a contextually-valued variable), and those that are not. If there is in fact experimental evidence that context-sensitive phenomena like evaluativity vary in the degree to which they are context-sensitive, this evidence seems to be *prima facie* incompatible with semantic theories of context-sensitivity.

Finally, the results reported here highlight the need for a careful, theoretically informed interpretation of experimental results that pays attention to the initial context of interpretation as well as to how the experimental stimuli themselves update this context. When we adopted the Bierwisch Test as the core of methodology in Experiments 1 and 2, we intended to avoid complications of context by presenting each stimulus in a null and thereby (presumably) neutral context. However, our results show that even if the discourse-initial null context was neutral, the actual context of interpretation for the stimuli ended up not being neutral because the descriptive content of the stimuli themselves is able to simultaneously draw

[20] We are grateful to an anonymous reviewer for this observation, and for mentioning the interesting alternative hypothesis that while evaluativity itself might not gradable, the likelihood of interpreting an adjective as evaluative could be – and this likelihood could be affected by a variety of factors: the salience of a potential comparison class, multidimensionality, relatedness to a clear measurement scale (e.g., *tall* vs. *easy*), frequency in general, and frequency of use in contexts with salient measurement scales in particular.

from and affect / update the context. Paying attention to the incremental dynamics of interpretation crucially enabled us to understand the fine-grained effects provided by our experimental studies, which were at first glance contradicting the coarser generalizations reported in the previous literature that were based on informally collected introspective judgments.

References

Armstrong, S., Gleitman, L., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13:263–308.

Bale, A. (2008). A universal scale of comparison. *Linguistics and Philosophy*, 31:1–55.

Barker, C. (2002). The dynamics of vagueness. *Linguistics and Philosophy*, 25:1–36.

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

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

Bierwisch, M. (1989). The semantics of gradation. In Bierwisch, M. and Lang, E., editors, *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*, pages 71–237. Springer-Verlag.

Bochnak, M. R. and Bogal-Allbritten, E. (2014). Investigating gradable predicates, comparison, and degree constructions in underrepresented languages. In Bochnak, M. R. and Matthewson, L., editors, *Methodologies in Semantic Fieldwork*. Oxford University Press.

Bogal-Allbritten, E. (2013). Decomposing notions of adjectival transitivity in Navajo. *Natural Language Semantics*, 21:277–314.

Bogal-Allbritten, E. (2015). Slightly coerced: Processing evidence for adjectival coercion by minimizers. In *Proceedings of the 48th Meeting of the Chicago Linguistic Society (CLS 48): Parasession on Meaning and Cognition*. Chicago: CLS Publications.

Breakstone, M. (2012). Inherent evaluativity. In Aguilar, A., Chernilovskaya, A., and Nouwen, R., editors, *Proceedings of Sinn und Bedeutung 16*. MITWPL.

Burnett, H. (2012). *The Grammar of Tolerance: On Vagueness, Context-Sensitivity, and the Origin of Scale Structure*. PhD Thesis, UCLA.

Cappelen, H. and Lepore, E. (2005). *Insensitive Semantics*. Wiley-Blackwell.

Christensen, R. H. B. (2012). ordinal—regression models for ordinal data. R package version 2012.09-11 http://www.cran.r-project.org/package=ordinal/.

Clark, H. (1969a). Influence of language on solving three-term series problems. *Journal of Experimental Psychology*, 82:205–215.

Clark, H. (1969b). Linguistic processes in deductive reasoning. *Psychological Review*, 76:387–404.

Clark, H. (1972). Difficulties people have answering the question, "Where is it?". *Journal of Verbal Learning and Verbal Behavior*, 11:265–277.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2012). Tolerant, classical, strict. *Journal of Logic*, 41:347–385.

Cresswell, M. (1976). The semantics of degree. In Partee, B., editor, *Montague Grammar*. Academic Press.

Cruse, D. A. (1976). Three classes of antonyms in English. *Lingua*, 38:281–292.

Cruse, D. A. (1980). Antonyms and gradable complementaries. In Kastovsky, D., editor, *Perspektiven der Lexikalischen Semantik: Beiträge zum Wuppertaler Semantikkolloquium vom 2Ű3, Dec. 1977*, pages 14–25. Bouvier.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Cummins, C. and Katsos, N. (2010). Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, 27:271–305.

Doetjes, J. (2012). On the (in)compatibility of non neutral adjectives and measure phrases. In Guevara, A., Chernilovskaya, A., and Nouwen, R., editors, *Proceedings of SuB 16*. MITWPL.

Doetjes, J., Constantinescu, C., and Součková, K. (2009). A neo-Kleinian approach to comparatives. In Cormany, E., Ito, S., and Lutz, D., editors, *Proceedings of SALT XIX*, pages 124–141. CLC Publications.

Donaldson, M. (1963). *A study of children's thinking*. Tavistock.

Doran, R., Baker, R., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, 1:211–248.

Flores D'Arcais, G. B. (1970). Linguistic structure and focus of comparison in processing of comparative sentences. In Flores D'Arcais, G. B. and Levelt, W. J. M., editors, *Advances in Psycholinguistics*. North-Holland.

Grano, T. (2012). Mandarin *hen* and universal markedness in gradable adjectives. *Natural Language and Linguistic Theory*, 30:513–565.

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

Haspelmath, M. and Buchholz, O. (1998). Equative and similative constructions in the languages of Europe. In van der Auwera, J., editor, *Adverbial constructions in the languages of Europe*, pages 277–334. Mouton de Gruyter.

Hellan, L. (1981). *Towards an Integrated Analysis of Comparatives*. Narr, Tübingen.

Higgins, E. T. (1977). The varying presuppositional nature of comparatives. *Journal of Psycholinguistic Research*, 6:203–222.

Horn, L. (1972). *On the Semantic Properties of the Logical Operators in English*. PhD thesis, UCLA.

Kamp, H. (1975). Two theories of adjectives. In Keenan, E., editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press.

Kamp, H. (1981). The paradox of the heap. In Mönnich, U., editor, *Aspects of philosophical logic*. Reidel.

Kamp, H. and Rossdeutscher, A. (1994). Drs-construction and lexically driven inferences. *Theoretical linguistics*, 20:165–235.

Katsos, N. (2008). The semantics/pragmatics interface from an experimental perspective: The case of scalar implicature. *Synthese*, 165:385–401.

Kennedy, C. (1999). *Projecting the Adjective: The syntax and semantics of gradability and comparison*. Garland Press.

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

Kennedy, C. (2009). Modes of comparison. In *CLS 43 Proceedings*, pages 141–165.

Kennedy, C. and Levin, B. (2008). Measure of change: the adjectival core of degree achievements. In McNally, L. and Kennedy, C., editors, *Adjectives and Adverbs: Syntax, Semantics and Discourse*. Oxford University Press.

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

Kim, C., Xiang, M., and Kennedy, C. (2013). Context dependence and shiftability in two classes of gradable adjectives. XPrag poster, Utrecht.

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

Klein, E. (1982). The interpretation of adjectival comparatives. *The Journal of Linguistics*, 18:113–136.

Kubota, Y. and Matsui, A. (2010). Modes of comparison and Question under Discussion: Evidence from 'contrastive comparison' in Japanese. In Li, N. and Lutz, D., editors, *Proceedings of SALT 20*, pages 57–75. CLC Publications.

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics*, 21:397–429.

McNabb, Y. (2012). Standard fixing and context manipulation: an experimental investigation of degree modification. In Aguilar-Guevara, A., Chernilovskaya, A., and Nouwen, R., editors, *Proceedings of Sinn und Bedeutung*, volume 16. MITWPL.

McNally, L. (2011). The relative role of property type and scale structure in explaining the behavior of gradable adjectives. *Lecture notes in computer science*, 6517:151–168.

Morzycki, M. (2009). Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. *Natural Language Semantics*, 17:175–207.

Morzycki, M. (2012). Adjectival extremeness: degree modification and contextually restricted scales. *Natural Language and Linguistic Theory*, 30:567–609.

Neeleman, A., van de Koot, H., and Doetjes, J. (2004). Degree expressions. *The Linguistic Review*, 21:1–66.

Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics*, 12(1):47–65.

Paradis, C. and Willners, C. (2006). Antonymy and negation – the boundedness hypothesis. *Journal of Pragmatics*, 38:1051–1080.

Piaget, J. (1928). *Judgment and reasoning in the child*. Paul Kegan.

Pinkal, M. (1989). Imprecise concepts and quantification. In Bartsch, R., van Benthem, J., and van Emde Boas, P., editors, *Semantics and contextual expression*, pages 221–265. Foris.

Potts, C. (2012). Conventional implicature and expressive content. In Maienborn, C., von Heusinger, K., and Portner, P., editors, *Semantics: an international handbook of natural language meaning*, volume 3, pages 2516–2536. Mouton de Gruyter.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rett, J. (2008a). Antonymy and evaluativity. In Gibson, M. and Friedman, T., editors, *Proceedings of SALT XVII*. CLC Publications.

Rett, J. (2008b). *Degree Modification in Natural Language*. PhD thesis, Rutgers University.

Rett, J. (2014). Modified numerals and measure phrase equatives. *Journal of Semantics*. online first, doi: 10.1093/jos/ffu004.

Rett, J. (2015). *The Semantics of Evaluativity*. Oxford University Press.

van Rooij, R. (2008). Comparatives and quantifiers. In Bonami, O. and Cabredo Hofherr, P., editors, *Empirical issues in formal syntax and semantics*, volume 7, pages 423–444. Centre national de la recherche scientifique (CNRS).

van Rooij, R. (2011). Implicit versus explicit comparatives. In Egré, P. and Klinedinst, N., editors, *Vagueness and language use*, volume 7, pages 51–72. New York: Palgrave MacMillan.

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

Sassoon, G. (2012a). A slightly modified economy principle: stable properties have nonstable standards. In Cohen, E., editor, *Proceedings of the Israel Association of Theoretical Linguistics (IATL)*, volume 27. MITWPL.

Sassoon, G. W. (2012b). A typology of multidimensional adjectives. *Journal of Semantics*, Advance Access Aug. 9 2012, doi:10.1093/jos/ffs012:1–46.

Sassoon, G. W. and Zevakhina, N. (2012). Granularity shifting: Experimental evidence from degree modifiers. In *Proceedings of SALT 22*. University of Chicago.

Sawada, O. (2009). Pragmatic aspects of implicit comparison: an economy-based approach. *Journal of Pragmatics*, 41:1079–1103.

Schwarzschild, R. (2012). Directed scale segments. In Chereches, A., editor, *Proceedings of SALT XXII*, pages 65–82. CLC Publications.

Seuren, P. (1984). The comparative revisited. *Journal of Semantics*, 3:109–141.

Solt, S. and Gotzner, N. (2012). Experimenting with degree. In Chereches, A., editor, *Proceedings of SALT 22*.

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, 3:1–77.

Sperber, D. and Wilson, D. (1986). *Relevance*. Basil Blackwell.

Stanley, J. and Szabo, Z. (2002). On quantifier domain restriction. *Mind and Language*, 15:219–61.

Stassen, L. (1985). *Comparison and Universal Grammar: an essay in universal grammar*. Basil Blackwell.

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

Toledo, A. and Sassoon, G. (2011). Absolute vs. relative adjectives – variance within vs. between indi-

viduals. In *Proceedings of SALT XXI*, pages 135–154. CLC Publications.
Tonhauser, J., Beaver, D., Roberts, C., and Simons, M. (2013). Toward a taxonomy of projective content. *Language*, 89:66–109.
Tribushinina, E. (2011). Piecemeal acquisition of boundedness. *Belgian Journal of Linguistics*, 25:93–116.
Tribushinina, E. and Gillis, S. (2012). The acquisition of scalar structures: production of adjectives and degree markers by Dutch-speaking children and their caregivers.
Whorf, B. L. (1956). The relation of habitual thought and behavior to language. In *Language, Thought, and Reality*, pages 134–159. MIT Press.
Yoon, Y. (1996). Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*, 4:217–236.
Zevakhina, N. and Geurts, B. (2011). Scalar diversity. Ms., National Research University at Nijmegen.

A Pilot experiment

The pilot experiment was designed as a broad initial foray into 3 dimensions of contrast: (*i*) construction type (positive, comparative, equative), (*ii*) adjective polarity (positive vs. negative polarity), and (*iii*) adjectival class (relative, absolute, total/partial). The main focus was on construction type and adjective polarity. Adjectival class was treated as a covariate, and not as an experimental manipulation. In contrast, when we pool together the data from Experiments 1 and 2 discussed in the main text, adjectival class is an explicit (between subjects) experimental manipulation. The pilot examined 28 adjective pairs from all 3 adjective classes, listed in Table 5 below according to their adjectival class:

	relative adjectives		**absolute adjectives**		**total/partial adjectives**	
	pos. pol.	*neg. pol*	*pos. pol.*	*neg. pol.*	*total/pos. pol.*	*partial/neg. pol.*
1	tall	short	complete	incomplete	clean	dirty
2	dark	light	full	empty	dry	wet
3	deep	shallow	opaque	transparent	flat	curved
4	even	uneven	open	closed	healthy	sick
5	expensive	inexpensive	perfect	imperfect	smooth	bumpy
6	fast	slow	visible	invisible	straight	bent
7	good	bad				
8	hard	easy				
9	heavy	light				
10	high	low				
11	kind	unkind				
12	large	small				
13	long	short				
14	strong	weak				
15	thick	thin				
16	wide	narrow				

Table 5: Pilot experiment – Adjective pairs

For each of these adjective pairs, five distinct constructions where examined: positive `pos` (the control / reference condition), comparatives with subject-targeted evaluativity `subjComp`, comparatives with object-targeted evaluativity `objComp`, equatives with subject-targeted evaluativity `subjEq`, and finally equatives with object-targeted evaluativity `objEq`. These are exemplified below with the `negPol` adjective *short* (from the `relAdj` class):

(40) a. `pos`: *Maria is short*
b. `subjComp`: *Maria is shorter than Sophie*
c. `objComp`: *Maria is shorter than Sophie*
d. `subjEq`: *Maria is as short as Sophie*
e. `objEq`: *Maria is as short as Sophie*

One item was constructed for each of the 28 `posPol-negpol` pairs above. Every participant saw each of these items twice, once with the `posPol` adjective and once with the `negPol` adjective, for a total of 56 stimuli (excluding fillers). Every stimulus item was in one of the 5 construction types listed above; the construction type was randomly selected for every one of the 56 stimuli and rotated for every participant (Latin square design). Thus, the two adjectives in a pair occurred in random order and with different constructions for every participant. And each participant was guaranteed to have (at least) 5 measurements with a `posPol` adjective and 5 with a `negPol` adjective for each of the 5 construction types. In addition to the 56 experimental stimuli, there were 58 fillers involving modal expressions. Thus, every participant responded to 114 stimuli (56 experimental stimuli + 58 fillers), the order of which was randomized for every participant. 117 participants (95 undergraduate students from UCSC and 22 undergraduate students from UCLA) completed the experiment online on a UCSC-hosted installation of the IBEX platform for course credit or extra-credit.

We examined the pattern of responses that the participants gave for the `pos` (positive) construction. This construction is the control and we expect an overwhelming amount of 2 or 1 responses for it from any given participant. If a participant has a consistent pattern of very low responses for the `pos` construction (relative to the majority of the other participants), it is likely that s/he did not complete the experiment properly and we should drop the data for that participant. 66 participants had mean responses of 2 (the absolute maximum), which is the normative, expected response behavior assuming absolutely no error on the part of the participants. The remaining 51 participants have mean responses less than 2 for the `pos` construction type, but only 11 of them have mean responses less than 1.25. Out of these 11 participants, 4 seem to have particularly conservative/cautious response patterns but they seem to have completed the experiment properly. We therefore only dropped the data of the 7 other participants.

We then examined the remaining 110 participants for repeated patterns of behavior that could not be detected by examining the low means for the `pos` construction. That is, we checked to see if any participants selected only 2 or 1 pretty much all the time throughout the experiment, which would strongly indicate that they did not complete the experiment properly. One participant has a clear extreme pattern of responses: 48 responses of 2 out of a total of 56. We dropped the data from this participant too. The final number of participants is 109 with 56 measurements per participant, for a total number of 6104 observations.

Figure 7 provides plots for all three adjective classes – `relAdj` in the first column from the left, `absAdj` in the second column and `tpAdj` in the third column – and for all five construction types – `pos` in the first row, `objComp` in the second row, `objEq` in the third row, `subjComp` in the fourth row, and finally `subjEq` in the fifth row.

Several generalizations emerge. First, entailment to the positive construction is clearly the most acceptable for the `pos` construction, while all the other constructions much less acceptably entail the positive construction. The differences between these constructions are fairly small, with `objComp` clearly the least acceptable entailment pattern. Second, there is no significant difference between `posPol` and `negPol` adjectives except for `tpAdj`, where the `posPol-negPol` constrast is confounded with the total-partial contrast. In general, total/`posPol` adjectives of the `tpAdj` class seem to behave more like `relAdj`, while partial/`negPol` adjectives seem to behave more like `absAdj`.

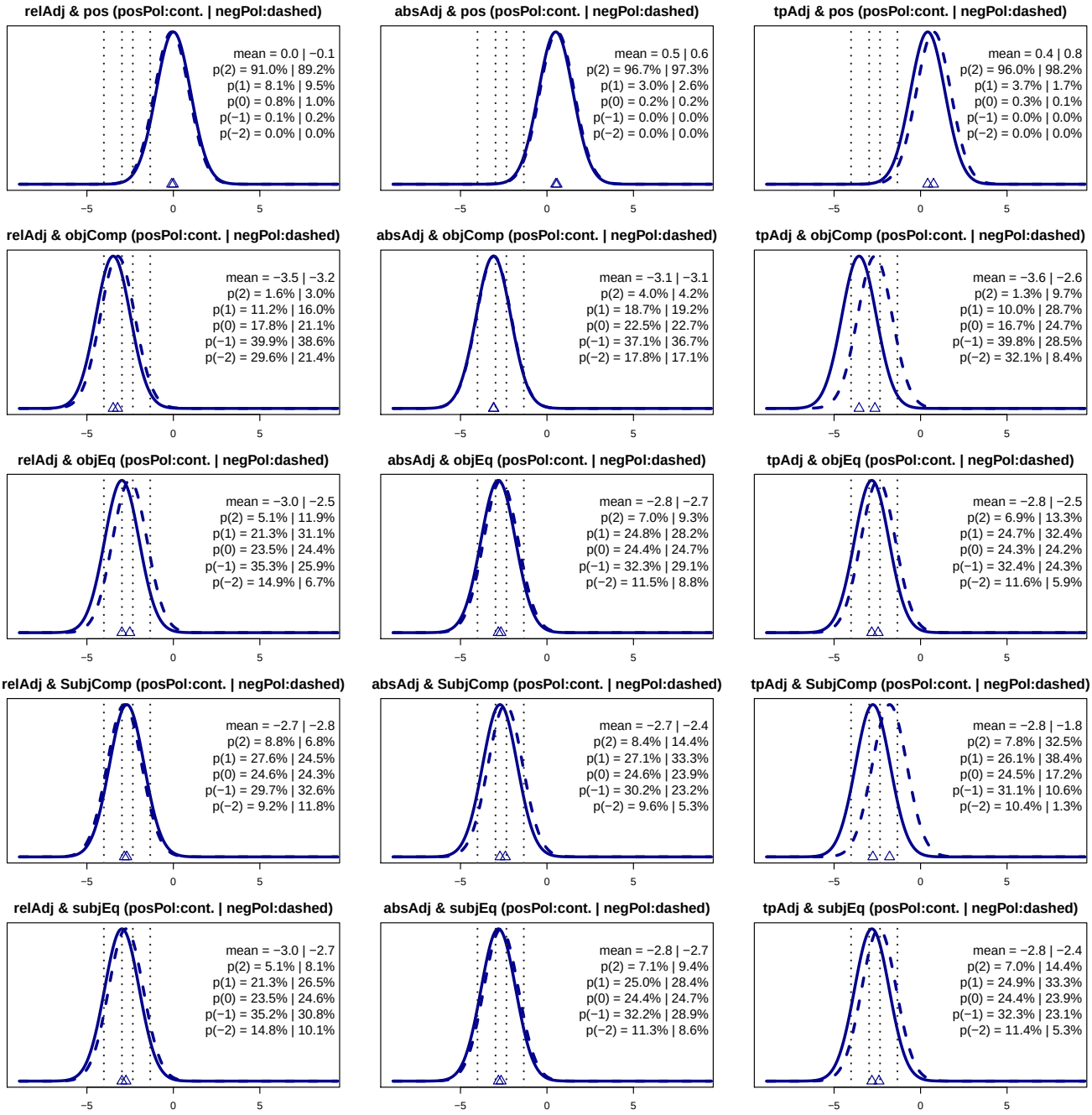

Figure 7: Pilot experiment – Plot of MLEs of the best mixed-effects ordinal probit model.

Fixed effects	Estimate	Std. Error	p-value
negPol	-0.10	0.12	
objComp	**-3.48**	0.11	0.00
objEq	**-2.98**	0.11	0.00
subjComp	**-2.69**	0.11	0.00
subjEq	**-2.97**	0.11	0.00
absAdj	**0.51**	0.22	0.02
tpAdj	**0.41**	0.21	0.05
negPol : objComp	**0.36**	0.15	0.02
negPol : objEq	**0.55**	0.15	0.0001
negPol : subjComp	-0.04	0.15	
negPol : subjEq	**0.33**	0.15	0.025
negPol : absAdj	0.18	0.29	
negPol : tpAdj	0.44	0.31	
objComp : absAdj	-0.12	0.23	
objEq : absAdj	-0.35	0.23	
subjComp : absAdj	**-0.53**	0.23	0.02
subjEq : absAdj	-0.34	0.23	
objComp : tpAdj	**-0.48**	0.23	0.035
objEq : tpAdj	-0.26	0.23	
subjComp : tpAdj	**-0.48**	0.23	0.04
subjEq : tpAdj	-0.25	0.23	
negPol : objComp : absAdj	-0.41	0.33	
negPol : objEq : absAdj	-0.48	0.33	
negPol : subjComp : absAdj	0.28	0.33	
negPol : subjEq : absAdj	-0.25	0.33	
negPol : objComp : tpAdj	0.22	0.35	
negPol : objEq : tpAdj	-0.52	0.34	
negPol : subjComp : tpAdj	**0.67**	0.35	0.05
negPol : subjEq : tpAdj	-0.26	0.34	
Thresholds	**Estimate**	**Std. Error**	
-2|-1	-4.02	0.12	
-1|0	-2.97	0.12	
0|1	-2.34	0.12	
1|2	-1.34	0.11	
Random effects	**Var**	**Std.Dev**	
subj	0.47	0.68	
item	0.02	0.15	

Table 6: Pilot experiment – MLEs for the best mixed-effects ordinal probit model (the reference level for adjective polarity is posPol, for adjective class is relAdj, and for construction type is pos)

Issues with the pilot experiment addressed in Experiments 1 and 2:

- adjective class was a covariate, i.e., it was not experimentally manipulated, so it is not clear how much we should conclude from the difference between relAdj and absAdj;
- every participant saw every one of the 28 items twice, once with the posPol adjective in an adjective pair, and once with the corresponding negPol adjective; while the two occurrences of the same item were always in different construction types (given our latin square design), this might have

dampened the effect of adjective polarity and might be the reason for the null effect of polarity in the pilot experiment;

- every adjective (pair) occurred in only one item, increasing the chance that the null adjective polarity effect or the observed difference between relative and absolute adjectives might be due to the items themselves rather than the adjectives under study; this issue is mitigated by the fact that we have item random effects in our ordinal probit model but is not eliminated.

B Experiments 1 and 2

B.1 Items, Experiment 1

For Experiments 1 and 2, we fully list the first item set and provide only the first two conditions for the remaining items.

(1) a. The Detective reported to the Police Chief: "Maria is tall."
The Chief concluded from this that Maria is tall.
b. Maria is as tall as Sophie. → Maria is tall.
c. Maria is as tall as Sophie. → Sophie is tall.
d. Maria is taller than Sophie. → Maria is tall.
e. Maria is taller than Sophie. → Sophie is tall.
f. Maria is short. → Maria is short.
g. Maria is as short as Sophie. → Maria is short.
h. Maria is as short as Sophie. → Sophie is short.
i. Maria is shorter than Sophie. → Maria is short.
j. Maria is shorter than Sophie. → Sophie is short.

(2) a. The oak sapling is tall. → the oak sapling is tall.
b. The oak sapling is as tall as the cedar sapling. → the oak sapling is tall.

(3) a. The zoo's koala is tall. → the zoo's koala is tall.
b. The zoo's koala is as tall as the zoo's bear cub. → the zoo's koala is tall.

(4) a. John's pea plant is tall. → John's pea plant is tall.
b. John's pea plant is as tall as his tomato plant. → John's pea plant is tall.

(5) a. Bill's cactus is tall. → Bill's cactus is tall.
b. Bill's cactus is as tall as Amy's cactus. → Bill's cactus is tall.

(6) a. The ceiling in the kitchen is tall. → the ceiling in the kitchen is tall.
b. The ceiling in the kitchen is as tall as the ceiling in the bathroom. → the ceiling in the kitchen is tall.

(7) a. The office's accountant is tall. → the office's accountant is tall.
b. The office's accountant is as tall as the office's secretary. → the office's accountant is tall.

(8) a. Jeremy's son is tall. → Jeremy's son is tall.
b. Jeremy's son is as tall as Dan's daughter. → Jeremy's son is tall.

(9) a. Peter's shirt is dark. → Peter's shirt is dark.
b. Peter's shirt is as dark as Hannah's shirt. → Peter's shirt is dark.

(10) a. The sky at dusk is dark. → the sky at dusk is dark.
b. The sky at dusk is as dark as it is at dawn. → the sky at dusk is dark.

(11) a. The sunglasses sold on the West Coast are dark. → the sunglasses sold on the West Coast are dark.
b. The sunglasses sold on the West Coast are as dark as those sold on the East Coast. → the sunglasses sold on the West Coast are dark.

(12) a. Wendy's skin is dark. → Wendy's skin is dark.
b. Wendy's skin is as dark as Bill's skin. → Wendy's skin is dark.

(13) a. Royal blue is dark. → royal blue is dark.
b. Royal blue is as dark as navy blue. → royal blue is dark.

(14) a. The television screen is dark. → the television screen is dark.
b. The television screen is as dark as the computer screen. → the television screen is dark.

(15) a. The ice in Greenland is dark. → the ice in Greenland is dark.
b. The ice in Greenland is as dark as the ice in Norway. → the ice in Greenland is dark.

(16) a. The pants are dark. → the pants are dark.
b. The pants are as dark now as they were before they were washed. → the pants are dark.

(17) a. The fee is expensive. → the fee is expensive.
b. The fee is as expensive as the tax. → the fee is expensive.

(18) a. Sarah's skateboard is expensive. → Sarah's skateboard is expensive.
b. Sarah's skateboard is as expensive as Will's skateboard. → Sarah's skateboard is expensive.

(19) a. A meat-lover's pizza is expensive. → a meat-lover's pizza is expensive.
b. A meat-lover's pizza is as expensive as a vegetarian pizza. → a meat-lover's pizza is expensive.

(20) a. Helen's tuition is expensive. → Helen's tuition is expensive.
b. Helen's tuition is as expensive as Betty's tuition. → Helen's tuition is expensive.

(21) a. The architecture class is expensive. → the architecture class is expensive.
b. The architecture class is as expensive as the business class. → the architecture class is expensive.

(22) a. The striped shirt is expensive. → the striped shirt is expensive.
b. The striped shirt is as expensive as the checked shirt. → the striped shirt is expensive.

(23) a. A bus fare is expensive. → a bus fare is expensive.
b. A bus fare is as expensive as a subway fare. → a bus fare is expensive.

(24) a. The bridge toll is expensive. → the bridge toll is expensive.
b. The bridge toll is as expensive as the tunnel toll. → the bridge toll is expensive.

(25) a. Linda's skates is fast. → Linda's skates is fast.
b. Linda's skates is as fast as Tom's skates. → Linda's skates is fast.

(26) a. Rodney's bike is fast. → Rodney's bike is fast.
b. Rodney's bike is as fast as Joe's bike. → Rodney's bike is fast.

(27) a. Jim's desktop computer is fast. → Jim's desktop computer is fast.
b. Jim's desktop computer is as fast as Tara's laptop computer. → Jim's desktop computer is fast.

(28) a. Waltzes are fast. → waltzes are fast.
b. Waltzes are as fast as sambas. → waltzes are fast.

(29) a. Professor Harris's lecture pace is fast. → Professor Harris's lecture pace is fast.
b. Professor Harris's lecture pace is as fast as Professor Wu's lecture pace. → Professor Harris's lecture pace is fast.

(30) a. The songs on the first album are fast. → the songs on the first album are fast.
b. The songs on the first album are as fast as the songs on the second album. → the songs on the first album are fast.

(31) a. Buses in Jim's hometown are fast. → buses in Jim's hometown are fast.
b. Buses in Jim's hometown are as fast as buses in Linda's hometown. → buses in Jim's hometown are fast.

(32) a. Desert wind is fast. → desert wind is fast.
b. Desert wind is as fast as sea wind. → desert wind is fast.

(33) a. Paul's test is hard. → Paul's test is hard.
b. Paul's test is as hard as Isabel's test. → Paul's test is hard.

(34) a. Meeting people online is hard these days. → meeting people online is hard these days.
b. Meeting people online is as hard as meeting people in person these days. → meeting people online is hard these days.

(35) a. Picking locks is hard these days. → picking locks is hard these days.
b. Picking locks is as hard as cracking safes. → picking locks is hard.

(36) a. California's driving exam is hard. → California's driving exam is hard.
b. California's driving exam is as hard as Washington's driving exam. → California's driving exam is hard.

(37) a. Meeting Gordon is hard. → meeting Gordon is hard.
b. Meeting Gordon is as hard as meeting Samantha. → meeting Gordon is hard.

(38) a. French is hard. → French is hard.
b. French is as hard as Spanish. → French is hard.

(39) a. Ballroom dancing is hard. → ballroom dancing is hard.
b. Ballroom dancing is as hard as hip-hop dancing. → ballroom dancing is hard.

(40) a. Living near the highway is hard. → living near the highway is hard.
b. Living near the highway is as hard as living near the train station. → living near the highway is hard.

(41) a. Bob's bike is heavy. → Bob's bike is heavy.
b. Bob's bike is as heavy as Juan's bike. → Bob's bike is heavy.

(42) a. Wood frames are heavy. → wood frames are heavy.
b. Wood frames are as heavy as plastic frames. → wood frames are heavy.

(43) a. Sandy's flower pots are heavy. → Sandy's flower pots are heavy.
b. Sandy's flower pots are as heavy as Tom's flower pots. → Sandy's flower pots are heavy.

(44) a. Jane's necklace is heavy. → Jane's necklace is heavy.
b. Jane's necklace is as heavy as Kim's necklace. → Jane's necklace is heavy.

(45) a. George's glass is heavy. → George's glass is heavy.
b. George's glass is as heavy as Valerie's glass. → George's glass is heavy.

(46) a. Mark's bag is heavy. → Mark's bag is heavy.
b. Mark's bag is as heavy as Julia's bag. → Mark's bag is heavy.

(47) a. The bowl of flour is heavy. → the bowl of flour is heavy.
b. The bowl of flour is as heavy as the bowl of sugar. → the bowl of flour is heavy.

(48) a. The box in Jenny's room is heavy. → the box in Jenny's room is heavy.
b. The box in Jenny's room is as heavy as the box in Josh's room. → the box in Jenny's room is heavy.

(49) a. Martha is strong. → Martha is strong.
b. Martha is as strong as Bertha. → Martha is strong.

(50) a. Hemp rope is strong. → hemp rope is strong.
b. Hemp rope is as strong as nylon rope. → hemp rope is strong.

(51) a. Canvas is strong. → canvas is strong.
b. Canvas is as strong as polyester. → canvas is strong.

(52) a. The students in Dorm A are strong. → the students in Dorm A are strong.
b. The students in Dorm A are as strong as the students in Dorm B. → the students in Dorm A are strong.

(53) a. Herman's grip is strong. → Herman's grip is strong.
b. Herman's grip is as strong as Sam's grip. → Herman's grip is strong.

(54) a. Bluejays are strong. → bluejays are strong.
b. Bluejays are as strong as robins. → bluejays are strong.

(55) a. The philosophy professor is strong. → the philosophy professor is strong.
b. The philosophy professor is as strong as its fullback. → the philosophy professor is strong.

(56) a. Encyclopedia bindings are strong. → encyclopedia bindings are strong.
b. Encyclopedia bindings are as strong as dictionary bindings. → encyclopedia bindings are strong.

(57) a. The city's sidewalk is wide. → the city's sidewalk is wide.
b. The city's sidewalk is as wide as the town's sidewalk. → the city's sidewalk is wide.

(58) a. The spots in Parking Lot A are wide. → the spots in Parking Lot A are wide.
b. The spots in Parking Lot A are as wide as the spots in Parking Lot B. → the spots in Parking Lot A are wide.

(59) a. The house's roof tiles are wide. → the house's roof tiles are wide.
b. The house's roof tiles are as wide as the garage's roof tiles. → the house's roof tiles are wide.

(60) a. The stripes on John's shirt are wide. → the stripes on John's shirt are wide.
b. The stripes on John's shirt are as wide as the stripes on his jacket. → the stripes on John's shirt are wide.

(61) a. The front steps are wide. → the front steps are wide.
b. The front steps are as wide as the back steps. → the front steps are wide.

(62) a. Joan's ladder is wide. → Joan's ladder is wide.
b. Joan's ladder is as wide as Luke's ladder. → Joan's ladder is wide.

(63) a. Tulip stems are wide. → tulip stems are wide.
b. Tulip stems are as wide as daisy stems. → tulip stems are wide.

(64) a. Modern watch bands are wide. → modern watch bands are wide.

b. Modern watch bands are as wide as traditional watch bands. → modern watch bands are wide.

B.2 Items, Experiment 2

(1) a. John's book is complete. → John's book is complete.
b. John's book is as complete as Bill's book. → John's book is complete.
c. John's book is as complete as Bill's book. → Bill's book is complete.
d. John's book is more complete than Bill's book. → John's book is complete.
e. John's book is more complete than Bill's book. → Bill's book is complete.
f. John's book is incomplete. → John's book is incomplete.
g. John's book is as incomplete as Bill's book. → John's book is incomplete.
h. John's book is as incomplete as Bill's book. → Bill's book is incomplete.
i. John's book is more incomplete than Bill's book. → John's book is incomplete.
j. John's book is more incomplete than Bill's book. → Bill's book is incomplete.

(2) a. Mary's instructions are complete. → Mary's instructions are complete.
b. Mary's instructions are as complete as Richard's instructions. → Mary's instructions are complete.

(3) a. Michael's symphony is complete. → Michael's symphony is complete.
b. Michael's symphony is as complete as Julie's symphony. → Michael's symphony is complete.

(4) a. The principal's report is complete. → the principal's report is complete.
b. The principal's report is as complete as the teacher's report. → the principal's report is complete.

(5) a. Yuki's dissertation is complete. → Yuki's dissertation is complete.
b. Yuki's dissertation is as complete as Doug's dissertation. → Yuki's dissertation is complete.

(6) a. The senate's bill is complete. → the senate's bill is complete.
b. The senate's bill is as complete as the congress' bill. → the senate's bill is complete.

(7) a. Abby's description of the problem is complete. → Abby's description of the problem is complete.
b. Abby's description of the problem is as complete as Libby's description of the problem. → Abby's description of the problem is complete.

(8) a. Ray's homework assignment is complete. → Ray's homework assignment is complete.
b. Ray's homework assignment is as complete as Maggie's homework assignment. → Ray's homework assignment is complete.

(9) a. The glass is full. → the glass is full.
b. The glass is as full as the mug. → the glass is full.

(10) a. The pool is full. → the pool is full.
b. The pool is as full as the jacuzzi. → the pool is full.

(11) a. The whiskey bottle are full. → the whiskey bottle are full.
b. The whiskey bottle are as full as the vodka bottle. → the whiskey bottle are full.

(12) a. George's locker is full. → George's locker is full.
b. George's locker is as full as Fred's locker. → George's locker is full.

(13) a. Gabby's bag is full. → Gabby's bag is full.
b. Gabby's bag is as full as Paul's bag. → Gabby's bag is full.

(14) a. The container of flour is full. → the container of flour is full.
b. The container of flour is as full as the container of sugar. → the container of flour is full.

(15) a. The perfume bottle is full. → the perfume bottle is full.
b. The perfume bottle is as full as the cologne bottle. → the perfume bottle is full.

(16) a. The pail is full. → the pail is full.
b. The pail is as full as the bucket. → the pail is full.

(17) a. The front partition is opaque. → the front partition is opaque.
b. The front partition is as opaque as the back partition. → the front partition is opaque.

(18) a. The writing paper is opaque. → the writing paper is opaque.
b. The writing paper is as opaque as the drawing paper. → the writing paper is opaque.

(19) a. A South American lizard's skin is opaque. → a South American lizard's skin is opaque.
b. A South American lizard's skin is as opaque as an African lizard's skin. → a South American lizard's skin is opaque.

(20) a. The pond water is opaque. → the pond water is opaque.
b. The pond water is as opaque as the lake water. → the pond water is opaque.

(21) a. The house's roof is opaque. → the house's roof is opaque.
b. The house's roof is as opaque as the house's back wall. → the house's roof is opaque.

(22) a. The water jug is opaque. → the water jug is opaque.
b. The water jug is as opaque as the wine jug. → the water jug is opaque.

(23) a. Ben's sunglasses are opaque. → Ben's sunglasses are opaque.
b. Ben's sunglasses are as opaque as Tina's sunglasses. → Ben's sunglasses are opaque.

(24) a. The new divider is opaque. → the new divider is opaque.
b. The new divider is as opaque as the old divider. → the new divider is opaque.

(25) a. The door is open. → the door is open.
b. The door is as open as the window. → the door is open.

(26) a. The red box is open. → the red box is open.
b. The red box is as open as the blue box. → the red box is open.

(27) a. The refrigerator door is open. → the refrigerator door is open.
b. The refrigerator door is as open as the freezer door. → the refrigerator door is open.

(28) a. The living room curtains are open. → the living room curtains are open.
b. The living room curtains are as open as the dining room curtains. → the living room curtains are open.

(29) a. The sleeping cat's mouth is open. → the sleeping cat's mouth is open.
b. The sleeping cat's mouth is as open as the sleeping dog's mouth. → the sleeping cat's mouth is open.

(30) a. The desk drawer is open. → the desk drawer is open.
b. The desk drawer is as open as the cabinet. → the desk drawer is open.

(31) a. The car's hood is open. → the car's hood is open.
b. The car's hood is as open as the car's trunk. → the car's hood is open.

(32) a. The blue folder is open. → the blue folder is open.
b. The blue folder is as open as the red folder. → the blue folder is open.

(33) a. Paul's test is perfect. → Paul's test is perfect.
b. Paul's test is as perfect as Isabel's test. → Paul's test is perfect.

(34) a. Sven's English is perfect. → Sven's English is perfect.
b. Sven's English is as perfect as Juana's English. → Sven's English is perfect.

(35) a. Brian's pronunciation is perfect. → Brian's pronunciation is perfect.
b. Brian's pronunciation is as perfect as Gail's pronunciation. → Brian's pronunciation is perfect.

(36) a. The paint job on the house is perfect. → the paint job on the house is perfect.
b. The paint job on the house is as perfect as the paint job on the shed. → the paint job on the house is perfect.

(37) a. The installation of the motor is perfect. → the installation of the motor is perfect.
b. The installation of the motor is as perfect as the installation of the battery. → the installation of the motor is perfect.

(38) a. Joe's recitation of the poem is perfect. → Joe's recitation of the poem is perfect.
b. Joe's recitation of the poem is as perfect as Danielle's recitation of the poem. → Joe's recitation of the poem is perfect.

(39) a. Josh's performance is perfect. → Josh's performance is perfect.
b. Josh's performance is as perfect as Gabe's performance. → Josh's performance is perfect.

(40) a. Valerie's knowledge of history is perfect. → Valerie's knowledge of history is perfect.
b. Valerie's knowledge of history is as perfect as Judy's knowledge of history. → Valerie's knowledge of history is perfect.

(41) a. The gas Robin discovered is visible. → the gas Robin discovered is visible.
b. The gas Robin discovered is as visible as the gas Mark discovered. → the gas Robin discovered is visible.

(42) a. Carol's blemishes are visible. → Carol's blemishes are visible.
b. Carol's blemishes are as visible as Henry's blemishes. → Carol's blemishes are visible.

(43) a. The effects of the first law of motion are visible. → the effects of the first law of motion are visible.
b. The effects of the first law of motion are as visible as the effects of the second law of motion. → the effects of the first law of motion are visible.

(44) a. The ghost in the story is visible. → the ghost in the story is visible.
b. The ghost in the story is as visible as the monster in the story. → the ghost in the story is visible.

(45) a. When applied, the cream is visible. → When applied, the cream is visible.
b. When applied, the cream is as visible as the gel. → When applied, the cream is visible.

(46) a. The effect of the air pollution is visible. → the effect of the air pollution is visible.
b. The effect of the air pollution is as visible as the effect of the water pollution. → the effect of the air pollution is visible.

(47) a. The zinc in the soil is visible. → the zinc in the soil is visible.
b. The zinc in the soil is as visible as the quartz in the soil. → the zinc in the soil is visible.

(48) a. The trace of the toxic chemical is visible. → the trace of the toxic chemical is visible.
b. The trace of the toxic chemical is as visible as the trace of the non-toxic chemical. → the trace of the toxic chemical is visible.

(49) a. The yellow wire is straight. → the yellow wire is straight.
b. The yellow wire is as straight as the red wire. → the yellow wire is straight.

(50) a. The steel rod is straight. → the steel rod is straight.
b. The steel rod is as straight as the iron rod. → the steel rod is straight.

(51) a. The downstairs railing is straight. → the downstairs railing is straight.
b. The downstairs railing is as straight as the upstairs railing. → the downstairs railing is straight.

(52) a. The train track is straight. → the train track is straight.
b. The train track is as straight as the subway track. → the train track is straight.

(53) a. Sue's metal bat is straight. → Sue's metal bat is straight.
b. Sue's metal bat is as straight as Lisa's metal bat. → Sue's metal bat is straight.

(54) a. Jill's hairpin is straight. → Jill's hairpin is straight.

(55) a. The short needle is straight. → the short needle is straight.
b. The short needle is as straight as the long needle. → the short needle is straight.

(56) a. The towel rack is straight. → the towel rack is straight.
b. The towel rack is as straight as the magazine rack. → the towel rack is straight.

(57) a. The baby is awake. → the baby is awake.
b. The baby is as awake as the babysitter. → the baby is awake.

(58) a. The student was awake. → the student was awake.
b. The student was as awake as the professor. → the student was awake.

(59) a. Jenny is awake. → Jenny is awake.
b. Jenny is as awake as Gabe. → Jenny is awake.

(60) a. Sarah's dog is awake. → Sarah's dog is awake.
b. Sarah's dog is as awake as Sarah's cat. → Sarah's dog is awake.

(61) a. The driver is awake. → the driver is awake.
b. The driver is as awake as the passengers. → the driver is awake.

(62) a. Matt was awake. → Matt was awake.
b. Matt was as awake as the priest. → Matt was awake.

(63) a. The secretary is awake. → the secretary is awake.
b. The secretary is as awake as the boss. → the secretary is awake.

(64) a. The security guard is awake. → the security guard is awake.
b. The security guard is as awake as the residents. → the security guard is awake.

B.3 Data summaries, Experiments 1 & 2

	rel. adj.: pos. pol.					rel. adj.: neg. pol.				
	objComp	objEq	pos	subjComp	subjEq	objComp	objEq	pos	subjComp	subjEq
2	14 (5%)	39 (14%)	229 (84%)	34 (13%)	41 (15%)	16 (6%)	47 (18%)	217 (81%)	29 (11%)	40 (7%)
1	54 (20%)	94 (35%)	36 (13%)	104 (38%)	90 (33%)	33 (12%)	90 (34%)	36 (13%)	97 (37%)	93 (22%)
0	56 (21%)	44 (16%)	6 (2%)	51 (19%)	58 (22%)	64 (24%)	51 (19%)	13 (5%)	50 (19%)	58 (22%)
-1	92 (34%)	66 (24%)	0 (0%)	68 (25%)	55 (20%)	100 (38%)	62 (23%)	3 (0%)	73 (28%)	58 (35%)
-2	53 (20%)	27 (10%)	1 (0%)	14 (5%)	26 (10%)	54 (20%)	17 (6%)	0 (0%)	16 (6%)	19 (15%)
	abs. adj.: pos. pol.					**abs. adj.: neg. pol.**				
	objComp	objEq	pos	subjComp	subjEq	objComp	objEq	pos	subjComp	subjEq
2	17 (6%)	22 (7%)	289 (93%)	43 (14%)	29 (9%)	28 (9%)	26 (9%)	288 (94%)	59 (19%)	24 (8%)
1	51 (17%)	53 (17%)	16 (5%)	76 (25%)	40 (13%)	50 (16%)	63 (21%)	13 (4%)	81 (27%)	63 (21%)
0	56 (18%)	75 (24%)	3 (1%)	67 (22%)	83 (27%)	61 (20%)	67 (22%)	3 (1%)	48 (16%)	69 (23%)
-1	103 (34%)	94 (30%)	2 (1%)	87 (28%)	90 (29%)	98 (32%)	87 (29%)	1 (0%)	73 (24%)	85 (28%)
-2	80 (26%)	66 (21%)	0 (0%)	34 (11%)	69 (22%)	70 (23%)	62 (20%)	1 (0%)	43 (14%)	64 (21%)

Table 7: Experiments 1 and 2 – Data summaries

B.4 Statistical modeling, Experiments 1 & 2

In the following model, the reference level for adjective polarity is `posPol`, for adjective class is `relAdj`, and for construction type is `pos`.

Fixed effects	**Estimate**	**Std. Error**	***p*-value**
`negPol`	**-0.20**	0.10	0.05
`objComp`	**-3.20**	0.11	0.00
`objEq`	**-2.44**	0.10	0.00
`subjComp`	**-2.42**	0.10	0.00
`subjEq`	**-2.40**	0.10	0.00
`absAdj`	**0.53**	0.20	0.008
`negPol : objComp`	0.16	0.12	
`negPol : objEq`	**0.26**	0.12	0.03
`negPol : subjComp`	0.14	0.12	
`negPol : subjEq`	0.19	0.12	
`negPol : absAdj`	**0.12**	0.06	0.04
`objComp : absAdj`	**-0.65**	0.12	0.00
`objEq : absAdj`	**-1.31**	0.12	0.00
`subjComp : absAdj`	**-0.77**	0.12	0.00
`subjEq : absAdj`	**-1.33**	0.12	0.00
Thresholds	**Estimate**	**Std. Error**	
`-2|-1`	-4.21	0.15	
`-1|0`	-3.05	0.15	
`0|1`	-2.32	0.15	
`1|2`	-1.24	0.15	
Random effects	**Var**	**Std.Dev**	
`subj`	0.63	0.79	
`item`	0.02	0.14	

Table 8: Experiments 1 and 2 – MLEs of the best mixed-effects ordinal probit model

The model with all 2-way interactions has been identified as the best one via a series of Likelihood ratio tests (LRTs); the random-effect structure for all models that we considered included intercept random effects for subjects and items:

- comparison with the main-effects only model: $p \approx 0, \chi^2 = 188.9, df = 9$
- comparison with the model that included the main effects and only the interaction between adjective polarity and construction type: $p \approx 0, \chi^2 = 182.6, df = 5$
- comparison with the model that included the main effects and only the interaction between adjective polarity and construction type on one hand and adjective polarity and adjective class on the other hand: $p \approx 0, \chi^2 = 177.9, df = 4$
- comparison with the model that included the main effects and only the interaction between adjective polarity and construction type on one hand and adjective class and construction type on the other hand: $p = 0.04, \chi^2 = 4.14, df = 1$
- comparison with the model that included the full effect structure, i.e., the main effects, all 2-way interactions as as all the 3-way interaction between adjective polarity, construction type and adjective class: $p = 0.21, \chi^2 = 5.83, df = 4$

C Experiment 3

C.1 Items, Experiment 3

For Experiment 3, we fully list the first item set and provide only the first condition for the remaining items.

(1) a. Maria is as tall as Sophie and Sophie is as tall as Joe. → Maria is tall.
b. Maria is as tall as Sophie and Sophie is as tall as Joe. → Sophie is tall.
c. Maria is taller than Sophie and Sophie is taller than Joe. → Maria is tall.
d. Maria is taller than Sophie and Sophie is taller than Joe. → Sophie is tall.
e. Maria is as short as Sophie and Sophie is as short as Joe. → Maria is short.
f. Maria is as short as Sophie and Sophie is as short as Joe. → Sophie is short.
g. Maria is shorter than Sophie and Sophie is shorter than Joe. → Maria is short.
h. Maria is shorter than Sophie and Sophie is shorter than Joe. → Sophie is short.

(2) a. The oak sapling is as tall as the cedar sapling and the cedar sapling is as tall as the pine sapling. → the oak sapling is tall.

(3) a. John's pea plant is as tall as his tomato plant and his tomato plant is as tall as his basil plant. → John's pea plant is tall.

(4) a. Bill's cactus is as tall as Amy's cactus and Amy's cactus is as tall as Tim's. → Bill's cactus is tall.

(5) a. Peter's shirt is as large as Hannah's shirt and Hannah's shirt is as large as Megan's. → Peter's shirt is large.

(6) a. Wendy's dog is as large as Bill's dog and Bill's dog is as large as Tim's. → Wendy's dog is large.

(7) a. The television screen is as large as the computer screen and the computer screen is as large as the tablet screen. → the television screen is large.

(8) a. The average house in Greenland is as large as the average house in Norway and the average house in Norway is as large as the average house in Germany. → the average house in Greenland is large.

(9) a. Sarah's skateboard is as expensive as Will's skateboard and Will's skateboard is as expensive as Tom's. → Sarah's skateboard is expensive.

(10) a. Helen's tuition is as expensive as Betty's tuition and Betty's tuition is as expensive as Jesse's. → Helen's tuition is expensive.

(11) a. The architecture class is as expensive as the business class and the business class is as expensive as the law one. → the architecture class is expensive.

(12) a. A bus fare is as expensive as a subway fare and a subway fare is as expensive as a train fare. → a bus fare is expensive.

(13) a. Linda's skates are as fast as Tom's skates and Tom's skates are as fast as Alex's. → Linda's skates are fast.

(14) a. Rodney's bike is as fast as Joe's bike and Joe's bike is as fast as Amy's. → Rodney's bike is fast.

(15) a. Jim's computer is as fast as Tara's computer and Tara's computer is as fast as Abby's. → Jim's computer is fast.

(16) a. Waltzes are as fast as sambas and sambas are as fast as flamencos. → waltzes are fast.

(17) a. Paul's test is as hard as Isabel's test and Isabel's test is as hard as Heather's. → Paul's test is hard.

(18) a. California's driving exam is as hard as Washington's driving exam and Washington's driving exam is as hard as Michigan's. → California's driving exam is hard.

(19) a. French is as hard as Spanish and Spanish is as hard as Italian. → French is hard.

(20) a. Ballroom dancing is as hard as hip-hop dancing and hip-hop dancing is as hard as the samba. → ballroom dancing is hard.

(21) a. Bob's bike is as heavy as Juan's bike and Juan's bike is as heavy as Mary's. → Bob's bike is heavy.

(22) a. Mark's bag is as heavy as Julia's bag and Julia's bag is as heavy as Carol's. → Mark's bag is heavy.

(23) a. The cereal bowl is as heavy as the soup bowl and the soup bowl is as heavy as the pasta bowl. → the cereal bowl is heavy.

(24) a. The box in Jenny's room is as heavy as the box in Josh's room and the box in Josh's room is as heavy as the one in Maria's room. → the box in Jenny's room is heavy.

(25) a. Martha is as strong as Bertha and Bertha is as strong as Paul. → Martha is strong.

(26) a. Hemp rope is as strong as nylon rope and nylon rope is as strong as plastic rope. → hemp rope is strong.

(27) a. Herman's grip is as strong as Sam's grip and Sam's grip is as strong as Maggie's. → Herman's grip is strong.

(28) a. Encyclopedia bindings are as strong as dictionary bindings and dictionary bindings are as strong as atlas bindings. → encyclopedia bindings are strong.

(29) a. The city's sidewalk is as wide as the town's sidewalk and the town's sidewalk is as wide as the village's. → the city's sidewalk is wide.

(30) a. The spots in Parking Lot A are as wide as the spots in Parking Lot B and the spots in Parking Lot B are as wide as the ones in Parking Lot C. → the spots in Parking Lot A are wide.

(31) a. The front steps are as wide as the back steps and the back steps are as wide as the basement steps. → the front steps are wide.

(32) a. Joan's ladder is as wide as Luke's ladder and Luke's ladder is as wide as Mia's. → Joan's ladder is wide.

C.2 Data summaries, Experiment 3

	pos. pol.				**neg. pol.**			
	subjComp1	subjComp2	subjEq1	subjEq2	subjComp1	subjComp2	subjEq1	subjEq2
2	40 (23%)	7 (4%)	9 (5%)	7 (4%)	41 (24%)	10 (6%)	15 (9%)	10 (6%)
1	76 (44%)	61 (35%)	65 (38%)	67 (39%)	82 (48%)	62 (36%)	59 (34%)	65 (38%)
0	27 (16%)	58 (34%)	50 (29%)	53 (31%)	17 (10%)	48 (28%)	48 (28%)	60 (35%)
-1	20 (12%)	32 (19%)	31 (18%)	33 (19%)	19 (11%)	42 (24%)	38 (22%)	25 (15%)
-2	9 (5%)	14 (8%)	17 (10%)	12 (7%)	13 (8%)	10 (6%)	12 (7%)	12 (7%)

Table 9: Experiment 3 – Data summaries

C.3 Statistical modeling, Experiment 3

In the following model, the reference level for adjective polarity is `posPol` and for construction type is `subjComp2`.

Fixed effects	Estimate	Std. Error	p-value
negPol	0.07	0.06	
subjComp1	**0.93**	0.089	0.00
subjEq1	0.05	0.085	
subjEq2	0.10	0.085	
Thresholds	**Estimate**	**Std. Error**	
-2|-1	-2.00	0.21	
-1|0	-0.70	0.19	
0|1	0.37	0.19	
1|2	2.19	0.20	
Random effects	**Var**	**Std.Dev**	
subj	1.36	1.17	
item	0.015	0.12	

Table 10: Experiment 3 – MLEs of the best mixed-effects ordinal probit model

The LRT comparing the main-effects only model (reported above) and the model with the full fixed-effect structure (i.e., including the 2-way interaction between adjective polarity and construction type); both models included intercept random effects for subjects and items: $p = 0.96, \chi^2 = 0.3, df = 3$.