Comparison classes and alternative salience in deriving adjectival upper bounds

Morwenna Hoeks (morwenna.hoeks@uni-osnabrueck.de)[1] **& Nicole Gotzner**[2]
[1-2]Institute of Cognitive Science, University of Osnabrück

Abstract

Listeners often strengthen the interpretation of expressions by deriving upper-bounded readings (URs), as when *warm* is understood as *warm but not hot*—an inference typically assumed to arise due to reasoning about a pragmatic scale ⟨*warm, hot*⟩. But, UR rates are also shown to vary substantially across and within scales (e.g., Aparicio and Ronai, 2025; Doran et al., 2009; Gotzner et al., 2018b; van Tiel et al., 2016). We investigate this diversity for gradable adjectives, aiming to explain why less URs arise for relative adjectives with relative scalemates (⟨*warm, hot*⟩) than absolute ones (⟨*likely, certain*⟩). We tease apart a view that attributes this effect to differences in salience of the stronger term from one attributing it to the way standards are set differently for these adjective types. Two experiments manipulated alternative salience and comparison classes while controlling for adjective type, showing that comparison classes modulate UR rates ⟨REL, REL⟩, but not ⟨REL, ABS⟩ scales, while alternative salience affects both. This pattern challenges salience-only accounts but is compatible with an account where URs depend on the way adjectival standards are derived from the comparison class.

Keywords: pragmatic inference; gradable adjectives; measurement scales; comparison classes

Introduction

Listeners often pragmatically strengthen the interpretation of linguistic expressions by drawing an upper bound on their meaning (Chierchia, 2013; Gazdar, 1979; Grice, 1975; Horn, 1972). For instance, hearers may infer from the water being described as *warm* that it is warm but not hot. In standard accounts, this strengthening takes place via reasoning about a pragmatic scale involving alternative linguistic expressions that convey a stronger meaning (e.g., Chierchia, 2013; Gazdar, 1979; Horn, 1972, a.m.o.): When a speaker states the water is *warm*, a hearer may infer that it is *warm but not hot* because if the speaker would assume the informationally-stronger term *hot* to hold, they should have stated it. Throughout, we use the notation ⟨*weak*, *strong*⟩ to refer to such pragmatic scales.

This literature tacitly assumes the mechanism for deriving these readings to be uniform across scales, but a growing body of evidence shows the likelihood for hearers to derive an upper bound varies significantly by pragmatic scale—a phenomenon called *scalar diversity* (Doran et al., 2009; Gotzner et al., 2018b; van Tiel et al., 2016). Recent evidence has also shown diversity *within* scales, with considerable variation in upper bounds when terms are embedded inside different carrier sentences (Aparicio & Ronai, 2023, 2025; Hu et al., 2022). The goal of this paper is to show that some of this variation can be explained by taking into account the semantic properties of expressions on these scales (following Aparicio and Ronai, 2025; Gotzner et al., 2018b; van Tiel et al., 2016).

Here, we focus specifically on upper bounded readings (URs) for gradable adjectives—adjectives like *warm, large* or *clean* that express gradable properties measured along a relevant dimension, such as temperature, size or cleanliness (Cresswell, 1976; Kennedy, 2007; Kennedy & McNally, 2005; Lassiter & Goodman, 2013; Solt, 2015). Crucially, gradable adjectives can be categorized into different semantic types which vary in their likelihood to trigger URs: Lower UR rates are typically found for *relative* adjectives (e.g., *warm but not hot, large but not gigantic*)—whose interpretation varies by context, than for *absolute* adjectives (e.g., *clean but not spotless*)—which exhibit a more context-invariant meaning (Doran et al., 2009; Gotzner et al., 2018b; van Tiel et al., 2016). For example, what temperature counts as *warm* depends on whether we are talking about a soda or coffee, but judgments about whether a cup is *clean* are much more consistent across these cases. Moreover, relative adjectives whose stronger scalemates are absolute, as in a pragmatic scale like ⟨*likely, certain*⟩ are also more likely to give rise to a UR (*likely but not certain*) than relative adjectives with alternatives that are also relative (e.g., ⟨*warm, hot*⟩, *warm but not hot*).

How these effects of adjective type on UR rates should be explained is less clear, however. These effects may be attributed to differences in the salience of the stronger term (Gotzner et al., 2018b; van Tiel et al., 2016), or, alternatively, may be related more directly to semantic differences among adjectives (Alexandropoulou et al., 2022; Aparicio & Ronai, 2023; Gotzner et al., 2018b). We thus aim to tease apart an account that attributes scalar diversity to differences in salience of stronger alternatives (ALTERNATIVE SALIENCE) from an account which seeks a more direct explanation in the way the meaning of relative and absolute adjectives depends differentially on context (MEASUREMENT; Gotzner, 2025). The two experiments reported below adjudicate between these accounts by orthogonally manipulating the contextual salience of the stronger scalemate and properties of the context that determine the semantics of such adjectives (i.e., the comparison class). Here, we focus specifically on comparing ⟨REL,REL⟩ with ⟨REL,ABS⟩ cases, since these pragmatic scales allow us to test the role of the stronger alternative while keeping the adjective type of the UR-trigger constant.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Two competing views for the scalar diversity effect

The ALTERNATIVE SALIENCE account explains the more robust inference patterns for absolute scalemates in terms of their relative salience: Absolute adjective scale may be more likely to come to mind in out-of-the-blue contexts than relative adjectives because the meaning of absolute adjectives involves an endpoint (see e.g., Gotzner et al., 2018b for such a suggestion). For instance, *certain*, *full* or *closed* all convey complete extents of the relevant properties.

This view follows (Neo-)Gricean accounts of URs in assuming that their derivation relies on the consideration of alternatives. In this view, URs come about via reasoning about alternative statements the speaker could have said, and such readings therefore only arise when hearers consider the possibility that the speaker would have uttered the alternative (Gazdar, 1979; Horn, 1972). Rees and Bott (2018) proposed a Salience Model, in which the derivation of an upper bound is initiated only once a stronger alternative exceeds a threshold of activation determined by discourse, semantic, and pragmatic factors. In such a model, differences in URs could thus be accounted for in terms of differences in activation of the relevant alternatives, with absolute adjectives, as end-point denoting terms, receiving more activation than relative ones.

In line with this view, van Tiel et al. (2016) found higher UR rates for pragmatic scales which included quantifiers and other expressions that denote an endpoint. Focusing on adjectival scales, Gotzner et al. (2018b) showed that a range of semantic properties account for the scalar diversity effect with pragmatic scales that involve absolute adjectives as the strong term yielding higher UR rates compared to pragmatic scales involving relative adjectives. What is more, mentioning stronger alternatives as part of the question under discussion leads to higher inference rates, particularly for adjectival scales. Doran et al. (2009). While such an effect is compatible with a role for the salience of alternatives, it may also indicate that alternatives need to be relevant as well as salient in order for URs to arise. And while Ronai and Xiang (2022) showed scalar diversity effects to be attenuated when part of a preceding question, they also showed such QUDs to not fully eliminate these effects.

Contrary to the ALTERNATIVE SALIENCE view, recent work has further indicated that the activation of alternatives is in fact a poor predictor of between-scale variability (Gotzner et al., 2018a; Hu et al., 2022; Lacina et al., 2024). For instance, van Tiel et al. (2016) results showed that alternative activation operationalized as cloze probability is a poor general predictor for UR rates. Moreover, Lacina et al. (2024), observed less priming for absolute than relative scalemates after the presentation of a gradable adjective prime, suggesting lower levels of activation for absolute adjectives. This finding speaks directly against the assumption that higher URs for absolute adjectives arise because such adjectives receive more activation than relative ones. We may therefore look instead more directly to semantic differences between relative and absolute adjectives to explain differential derivation of URs.

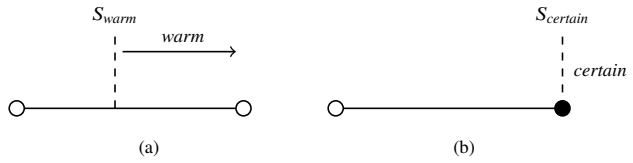

Figure 1: Standards of comparison (S) for (a) relative and (b) maximum absolute adjectives.

A second potential explanation for the scalar diversity effect, discussed for instance in Gotzner (2025) and van Tiel et al. (2016), is that differences in UR rates for relative and absolute adjectives stem from the fact that the meaning of two scalemates may be less easily distinguishable from one another when both are relative. To better understand such an account, we introduce some background on the semantics of gradable adjectives below.

In the literature on gradable adjectives, they are taken to map their arguments onto a measurement scale—a partially ordered set of degrees ordered along some dimension like temperature, likelihood or aperture—on which they set a lower bound (Kennedy, 2007; Kennedy & McNally, 2005). Accordingly, an object can be considered warm iff its temperature exceeds a relevant standard of comparison. This is depicted in Figure 1, where S_{warm} indicates the standard for *warm*. We can thus distinguish two notions of scales: a *measurement* scale on which gradable adjectives set lower (and potentially upper) bounds, and a *pragmatic* scale consisting of alternative expressions, ordered based on informativity.

As outlined above, one crucial difference between relative and absolute adjectives is that relative adjectives are associated with fully open scales that lack endpoints, while absolute adjectives employ scales that are closed at least on one end (indicated with closed circles in Figure 1). And while the standards for absolute adjectives are fixed at one of these endpoints, standards for relative adjectives are contextually determined (Kennedy, 2007; Kennedy & McNally, 2005; Solt, 2015). Specifically, relative standards depend on an inferred class of objects against which the object in question is compared, also called a *comparison class* (Bale, 2011; Lassiter & Goodman, 2013; Solt, 2009, 2015; Solt & Gotzner, 2012). For example, with a comparison class consisting of other cups of coffee, the minimum temperature for an object to be considered *warm* is higher than with a comparison class of sodas.

Because of this context dependence, listeners may be uncertain as to where the standard for each term on a ⟨REL, REL⟩ scale is located, resulting in less easily distinguishable meanings for these terms (see also (Gotzner et al., 2018a; Leffel et al., 2019)). Standards for absolute adjectives, in contrast, are anchored at a scalar endpoint and are clearly distinct from that of their scalemate regardless of their context of use. Higher rates of URs may therefore be observed for ⟨REL, ABS⟩ scales because their meanings can be mapped onto distinct ranges of degrees on the underlying measurement scale.

This view is compatible with empirical findings showing

that context-driven expectations for the scale under consideration predict the scalar diversity effects rather than explicit mention of stronger alternatives (Hu et al., 2022). It is moreover consistent with studies showing that comparison classes not only affect lower but also upper bounds for gradable adjectives, especially relative ones (Alexandropoulou et al., 2022; Aparicio & Ronai, 2025). This work on the role of comparison classes in reasoning about lower and upper bounds motivated the current study and is discussed in the following.

Empirical background on comparison classes in setting lower and upper bounds

Supporting the theoretical differences between relative and absolute adjectives, empirical work has shown that comparison classes affects judgments of the lower bound for relative but not absolute adjectives (e.g., Barner and Snedeker, 2008; Schmidt, 2009; Solt and Gotzner, 2012; Weicker and Schulz, 2024). More recent work has also shown that comparison classes directly affect adjectival upper bounds. For instance, Alexandropoulou et al. (2022) showed that the incremental computation of URs was facilitated for relative adjectives when the visually depicted comparison class contained a contrast object (representing an antonymic meaning), resulting in the respective standards to be more easily distinguishable. The more general finding that UR-rates are positively correlated with the difference in strength between scalemates (van Tiel et al., 2016) is also consistent with the idea that URs are boosted with higher distinguishability of the involved expressions.

Aparicio and Ronai (2023) argue, more specifically, that distance between adjectival standards drives variation in UR rates, even among instances of the same pragmatic scale. Instead of measuring how easily distinguishable the weak and strong term were from one another, as in van Tiel et al. (2016), they showed that distance between adjectival standards varies with the subject noun that the adjectives are predicated of, with nouns that are biased towards one end of the measurement scale in terms of their expected degree (e.g., *the coffee is warm*) generally involving smaller distances than neutral ones (e.g., *the water is warm*). If these nouns determine the relevant comparison class against which the adjective is interpreted, then neutral nouns would instantiate a wider range of degrees. As a result, adjectives with neutral nouns involve a larger distance between the relevant adjectival standards, boosting UR-rates in such sentences.

However, Aparicio and Ronai (2025) did not manipulate or control for the type of pragmatic scale, and it is thus unclear how the distance effect interacts with adjective type. If their observed effect of threshold distance is indeed driven by an underlying notion of threshold distinguishability, as suggested above, we would expect the manipulation of noun bias to have a larger impact on UR rates for ⟨REL, REL⟩ than for ⟨REL, ABS⟩ scales. This is because even though the standard for the weak, relative adjective may be dependent on the comparison class, it would still be easily distinguishable from the standard of the absolute stronger alternative that is located at a scalar endpoint. This is depicted in Figure 2, where standards for weak and strong adjectives on both types of scales are depicted with large (a-b) and small (c-d) threshold distances.[1]

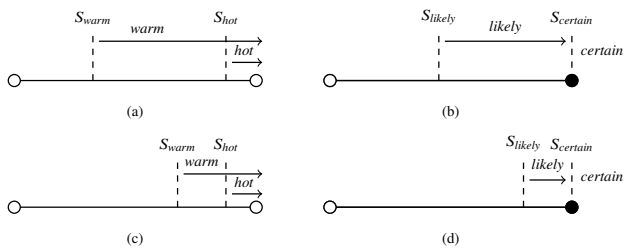

Figure 2: ⟨REL, REL⟩ (a and c) and ⟨REL, ABS⟩ scales (b and d) with varying standards of comparison (S) per scalemate.

Moreover, Aparicio and Ronai (2025) only tested UR rates without encouraging reasoning about scalemates, so it is unclear how such effects are modulated by reasoning about the salience of alternatives. The present work addresses these gaps by examining how comparison classes and alternative salience jointly shape the derivation of URs for ⟨REL, REL⟩ and ⟨REL, ABS⟩ scales. UR rates for ⟨REL, REL⟩ scales were tested in Experiment 1, and those for ⟨REL, ABS⟩ scales were tested in Experiment 2. Both Experiment 1 and 2 used contexts in which the stronger scalemate was either made salient or not, and both experiments used noun bias as in Aparicio and Ronai (2025) to manipulate the comparison class.

Hypotheses

In short, while the ALTERNATIVE SALIENCE view explains the connection between UR rates and adjective type in terms of inherent differences in salience of the stronger alternative, the MEASUREMENT view suggests URs to arise from a combination of the degree-based semantics of adjectives and pragmatic considerations about the comparison class determining the adjectival standards. These views make distinct predictions about how URs should vary as a function of scale type, scalemate salience, and comparison class properties.

The ALTERNATIVE SALIENCE view predicts that increasing the salience of the stronger alternative—e.g., by mentioning it in the question under discussion—boosts UR rates across scale types. However, this effect may be stronger for ⟨REL, REL⟩ scales, where alternatives are less salient and are therefore more likely to fall below the activation threshold required for UR derivation in the absence of contextual support. Crucially, because this account locates the source of scalar diversity in alternative activation rather than in comparison class representations, it predicts no systematic effect noun bias—unless this manipulation also interacts with the activation of alternatives. One possibility, explored in Hu et al. (2022), is that the subject noun affects predictability of the stronger

[1]For ease of presentation we depict *likely* and *certain* on one measurement scale that is closed on one end, but note that another possibility is that these adjectives are associated with entirely separate scales (Kennedy, 2007).

scalemate. In this vein, it could also be that biased nouns instantiating high degrees on the relevant scale involve greater predictability of the stronger alternative (e.g., mention of *coffee* may make the term *hot* more predictable). This hypothesis is therefore also compatible with an effect of noun bias where biased nouns give rise to more URs than neutral ones.

In contrast, the MEASUREMENT view predicts that UR rates depend on how clearly the standards for the weak and strong alternative can be distinguished on the underlying measurement scale. For ⟨REL, REL⟩ scales, where both adjectives involve comparison-class–dependent standards, URs should be sensitive to manipulations that affect threshold placement. Specifically, noun bias should modulate UR rates in that case: neutral nouns that give rise to broader distributions of degrees in the comparison classes and greater threshold distance should yield more URs than biased nouns that compress the relevant range of degrees, as found in Aparicio and Ronai (2025). For ⟨REL, ABS⟩ scales, in contrast, threshold distinguishability is expected to be high regardless of the comparison class, because the absolute adjective is anchored to a scalar endpoint (see Figure 2). As a result, UR rates for these scales should be less sensitive, or entirely insensitive, to noun bias. For both ⟨REL, REL⟩ and ⟨REL, ABS⟩ scales, however, increasing scalemate salience may still boost UR rates by facilitating pragmatic reasoning about salient standards, so this hypothesis does not rule out any additional effects of contextual salience of alternatives.

Taken together, the MEASUREMENT hypothesis uniquely predicts that comparison class manipulations affect UR rates for ⟨REL, REL⟩ scales (Exp1) but not for ⟨REL, ABS⟩ scales (Exp2). This pattern allows us to empirically distinguish threshold-based from salience-only accounts of scalar diversity.

Methods

The experiments were designed to adjudicate between the two views introduced above by manipulating alternative salience and comparison class properties across different types of pragmatic scales. Both experiments adopted a 2x2 factorial design crossing Noun Bias (NEUTRAL VS BIASED) with Question Type—manipulating whether or not the stronger alternative was made salient in the preceding question (ALT) or not (NO ALT). In addition, Adjective Type was included as a between-subjects manipulation, testing ⟨REL, REL⟩ scales in Experiment 1 and ⟨REL, ABS⟩ scales in Experiment 2.

Materials

All stimuli were presented in a short dialogue format, with a target sentence involving a relative adjective preceded by a question (see Figure 3). The experimental design of both experiments crossed Question Type (ALT VS. NO ALT) with Noun Bias (NEUTRAL VS. BIASED). In the NO ALT conditions of both experiments, participants were first asked a general question that did not mention the stronger scalemate (e.g., *What was the water like?*), followed by a reply containing the weak adjective. In the ALT condition, the initial question explicitly mentioned a stronger alternative, which was always an adjective that set a higher standard on the relevant scale (e.g., *Was the water hot?*), thereby increasing its salience. This stronger scalemate was always a relative adjective in Experiment 1 and an absolute adjective in Experiment 2.

To implement the Noun Bias manipulation, we adopted a subset of Aparicio and Ronai's (2023) pre-normed stimuli, but split them by adjective type. To rule out any explanation of comparison class effects in terms of absolute differences in the possible degrees between the comparison classes, we also excluded stimuli involving nouns with non-overlapping distributions of degrees (e.g., *large beach/elephant*). The resulting set of 28 ⟨REL, REL⟩ adjective pairs were then supplemented with 22 additional ones classified as ⟨REL, REL⟩ in the scalar diversity literature (Gotzner et al., 2018b; van Tiel et al., 2016), resulting in a total of 50 adjective pairs for Experiment 1. For Experiment 2, 11 ⟨REL, ABS⟩ adjective pairs from Aparicio and Ronai (2025) were supplemented with an additional set of 29 ⟨REL, ABS⟩ scales, resulting in a total of 40 adjective pairs for Experiment 2. Such pairs were selected only if the strong term passed the standard tests for absolute adjectives, such as compatibility with modifiers like *completely, fully or 100%* (see the OSF repository for a summary of these tests). Although noun bias was not independently normed for the entire stimulus set, for all newly constructed adjective pairs, NEUTRAL and BIASED subject nouns were created based on Aparicio and Ronai's manipulation of noun bias, with BIASED nouns always comprising objects or individuals instantiating high degrees on the underlying measurement scale, and NEUTRAL nouns involving a more neutral distribution of degrees.

Stimuli of each experiment were divided over four lists using a Latin square design. To counterbalance for response type and to check participants' attention, 45 fillers were interspersed with the target stimuli, involving adjective pairs with mutually inconsistent and unrelated meanings, as well as reversals of weak and strong adjectives inside questions and answers. As a pilot for Experiment 2, Experiment 1 included 10 ⟨REL, ABS⟩ adjective pairs from Aparicio and Ronai (2025). Exp2 included 10 ⟨REL, REL⟩ adjective pairs from Exp1.

Participants

A total of 200 native speakers of English were recruited online via the Prolific platform, 100 per experiment. Participants were located in the United States or Canada, and had an average approval rate of 85%. All participants provided written consent before taking part in the experiment and were rewarded at a $12 hourly rate. Participants of Experiment 1 were excluded from participation in Experiment 2.

Procedure

After reading each conversation, participants were asked to judge whether or not the answer gave rise to the UR to the negation of the stronger scalemate, as shown in Figure 3. Before presentation of the experimental trials, participants went through a brief practice phase familiarizing them with

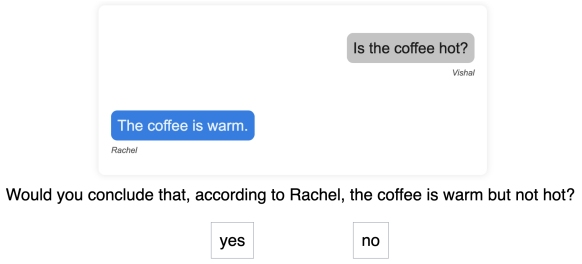

Figure 3: Example stimulus

Table 1: Posterior estimates for Experiment 1

	Estimate	Error	95%CrI	Rhat
Intercept	1.72	0.22	[1.29, 2.16]	1.00
Noun Bias	0.82	0.19	[0.46, 1.19]	1.00
Quest. Type	-0.36	0.14	[-0.65, -0.11]	1.00
NB:QT	0.14	0.20	[-0.25, 0.53]	1.00

the task. Both experiments lasted around 20 minutes.

Preregistrations and Data Availability

Hypotheses were preregistered on the Open Science Framework. Our experimental stimuli as well as the data and analysis scripts are also freely available there (link to OSF).

Results

Average UR rates for both experiments are shown with standard errors in Figure 4. From Experiment 1, one participant was removed due to not meeting language requirements, and an additional four participants were removed because their accuracy on attention checks fell below 60%. Nine participants from Experiment 2 were excluded due to failure on attention checks. Data from the remaining 186 participants were analyzed using R, version 3.6.3 (R Core Team, 2022). Separate Bayesian logistic linear mixed-effect models were fit to the response data from Experiment 1 and 2 using Stan, as implemented in the brms package, version 2.18.0 (Bürkner, 2017), with the default (flat) priors. The model included fixed effects for Noun Bias and Question Type (both coded -0.5/0.5) and a maximal random effects structure. Both models were ran for four chains, each with 4000 steps (warmup = 1000 steps). Rhat statistics approached 1.00 and no warnings emerged.

Posterior estimates of the model for Experiment 1 are shown in Table 1. This model revealed a main effect of Noun Bias, indicating higher UR rates in NEUTRAL than BIASED conditions. The model also revealed a main effect of Question Type, suggesting more URs when the stronger alternative was made salient in the preceding question. The 95% credible interval for the interaction term overlapped with zero indicating no reliable interaction effect.

The model again revealed a reliable main effect of Question Type, indicating higher UR rates when the stronger absolute adjective was salient in the question (see Table 2 for the full model outputs). Unlike Experiment 1, there was no reliable main effect of Noun Bias. There was also no evidence for an interaction between Noun Bias and Question Type. Pairwise comparisons confirmed no reliable differences between neutral and biased nouns in either Question Type condition.

Table 2: Posterior estimates for Experiment 2

	Estimate	Error	95%CrI	Rhat
Intercept	1.58	0.19	[1.21, 1.96]	1.00
Noun Bias	0.28	0.13	[0.02, 0.55]	1.00
Quest. Type	-0.04	0.11	[-0.25, 0.17]	1.00
NB:QT	0.25	0.23	[-0.19, 0.70]	1.00

To test the hypotheses using Bayes factors, the models described above were fit with weakly informative priors ($\mathcal{N}(0, 0.3)$, reflecting an expectation of small-to-moderate effects centered on zero). Bayes factors were computed using the Savage–Dickey density ratio. The model for Experiment 1 revealed very strong evidence for a main effect of Question Type (BF_{10} = 498.65) and strong evidence for a main effect of Noun Bias (BF_{10} = 11.62). The interaction was not supported (BF_{10} = 0.71), providing only weak anecdotal evidence in favor of the null hypothesis. The model for Experiment 2 revealed moderate evidence for a main effect of Question Type (BF_{10} = 3.32), but moderate evidence for the *null hypothesis* regarding the main effect of Noun Bias (BF_{10} = 0.36). Again, the model provided only weak anecdotal evidence in favor of the null for the interaction (BF_{10} = 0.81).

General Discussion

This paper investigated how comparison classes and alternative salience jointly shape the derivation of upper-bounded readings (URs) for gradable adjectives. Across two experiments, we examined the effects of alternative salience and comparison class properties on UR endorsement rates for ⟨REL, REL⟩ and ⟨REL, ABS⟩ pragmatic scales. The results revealed that, while alternative salience exerted an effect across scale types, comparison class manipulations affected UR rates for ⟨REL, REL⟩ scales, but not for ⟨REL, ABS⟩ scales.

This asymmetry supports the hypothesis that the comparison class plays a role in UR derivation only when both scalemates involve context-dependent thresholds, as suggested by the MEASUREMENT view. Under this view, differences in UR rates within and between adjectival scales arise from reasoning about the degree-based semantics of adjectives. When the standards of the weak and strong terms are clearly distinguishable—as is the case if one of them is end-point denoting—URs are more robust to variation in the comparison class. In contrast, when weak and strong adjectives rely on comparison-class–dependent standards, as in ⟨REL, REL⟩ scales, uncertainty about threshold placement can hinder UR derivation, particularly when the comparison class instantiates a narrow range of degrees, with a smaller distance between adjectival thresholds. This interpretation of the results also aligns with previous work showing that UR rates correlate with semantic distance (van Tiel et al., 2016) and with exper-

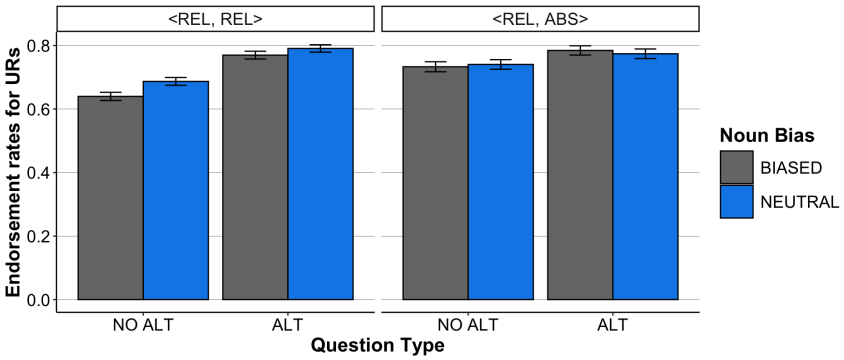

Figure 4: Mean endorsement rate for the upper-bounded reading per condition for Experiment 1 and Experiment 2

imental evidence that comparison classes instantiating more distinct degrees facilitate UR derivation specifically for ⟨REL, REL⟩ scales (Alexandropoulou et al., 2022).[2]

Within the MEASUREMENT view, the additional effect of alternative salience is compatible with at least two implementations. On one implementation, participants still engage in explicit reasoning about alternative terms, with URs arising only when (i) a relevant alternative exceeds an activation threshold and (ii) its standard is sufficiently distinguishable from that of the weak term. In this version, low threshold distinguishability can prevent UR derivation even when the alternative is highly salient, in line with the idea proposed in Leffel et al. (2019) that UR derivation is blocked whenever a borderline contradiction could be derived. This could also account for the negative correlation between priming and UR rates reported by Lacina et al. (2024): relative adjective alternatives may become strongly activated during comprehension of a weaker prime, yet fail to trigger a UR in some cases where the relevant thresholds are insufficiently distinct.

A second possible implementation, as suggested Gotzner (2025), is that upper bounds for gradable adjectives do not come about via (Neo-)Gricean reasoning about alternative expressions, but instead involve reasoning about the underlying measurement scale directly. In this view, URs are part of the meaning of these adjectives, which involve a range of degrees whose lower and upper bounds both depend on the derived comparison class. In that case, the effects of alternative salience emerge only indirectly: higher salience of the stronger term increases accessibility of the corresponding adjectival standard, which in turn shapes participants' judgments about the meaning of the weak term. Such an implementation would predict alternative salience to matter most for gradable adjectives, because reasoning about adjectival standards would only be involved for these expressions. Suggestive evidence for this comes from Doran et al.'s (2009), who found the inclusion of stronger scalemates in context questions to only boost UR rates for pragmatic scales with gradable adjectives, but not for quantifiers or cardinals. Future work should distinguish between these two potential implementations.

Surprisingly, UR endorsement rates were relatively high in both experiments. A possible contributor is the experimental setup, in which the UR-trigger always occurred in chat format, inside the answer to a preceding question. Since salience was moreover manipulated within-subjects, a stronger scalemate appeared in half of the trials, which may have signaled reasoning about alternatives to generally be relevant to the task. While these design choices were intentional and necessary to test salience-based predictions, they may also have increased participants' general readiness to compute URs, even in conditions where the alternative was not explicitly introduced. The inclusion of absolute adjective stimuli may similarly have boosted UR-rates in Experiment 1. Future work could mitigate this by adopting a between-subjects salience manipulation.

Some limitations of our experiments should also be acknowledged. First, while we relied on previously validated materials to classify nouns and adjectives, future work would benefit from explicit norming of the full stimuli set. Such norming would allow for a more fine-grained assessment of how gradience in threshold distance relates to UR rates across and within scales. Moreover, future studies should also test ⟨REL, REL⟩ and ⟨REL, ABS⟩ scales within the same experimental paradigm to more directly assess by-participant interactions between scale structure and contextual manipulations.

Despite these limitations, the central pattern of results is robust: comparison class manipulations selectively affected UR rates for ⟨REL, REL⟩ scales, with no reliable effects for ⟨REL, ABS⟩ ones. This selective sensitivity is difficult to reconcile with purely salience-based accounts of scalar diversity. Instead, the findings support a view in which implicature derivation at least also depends on the distinguishability of the relevant standards on the underlying measurement scale. More broadly, the present work contributes to a growing literature emphasizing the role of uncertainty and degree-based reasoning in the pragmatic processes responsible for deriving upper bounded interpretations.

[2]These results also speak against an alternative account of the noun bias effect, proposed in Aparicio and Ronai (2025), according to which biased nouns reduce URs because the strong term is a priori more likely to hold (e.g., coffee is more likely to be hot, making participants less willing to negate it). Crucially, biased nouns also increase prior likelihood of strong terms for ⟨REL, ABS⟩ scales, so this account would predict a noun bias effect independent of scale type, contrary to our findings.

Acknowledgments

This research was funded by a Research Fellowship of the Alexander von Humboldt Foundation awarded to the first author. NG is further supported by the DFG (Emmy Noether grant, project number 441607011).

References

Alexandropoulou, S., Herb, M., Discher, H., & Gotzner, N. (2022). Incremental pragmatic interpretation of gradable adjectives: The role of standards of comparison. *Semantics and linguistic theory*, 481–497.

Aparicio, H., & Ronai, E. (2023). Scalar implicature rates vary within and across adjectival scales. *Semantics and linguistic theory*, 110–130.

Aparicio, H., & Ronai, E. (2025). Scalar implicature rates vary within and across adjectival scales. *Journal of Semantics*, *42*(1-2), 97–126.

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, *19*(2), 169–190.

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

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*, 1–28.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP Oxford.

Cresswell, M. J. (1976). The semantics of degree. In B. H. Partee (Ed.), *Montague grammar* (pp. 261–292). Academic Press. https://doi.org/10.1016/B978-0-12-545850-4.50015-7

Doran, R., Baker, R. E., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, *1*, 1–38.

Gazdar, G. (1979). *Pragmatics, implicature, presuposition and logical form*. Academic Press.

Gotzner, N. (2025). The measurement mechanism: A new theory of the (lower- and) upper-bounded meaning of gradable adjectives. https://doi.org/10.13140/RG.2.2.33509.92645/1

Gotzner, N., Solt, S., & Benz, A. (2018a). Adjectival scales and three types of implicature. *Semantics and linguistic theory*, 409–432.

Gotzner, N., Solt, S., & Benz, A. (2018b). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, *9*, 1659.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Horn, L. R. (1972). *On the semantic properties of logical operators in english*. University of California, Los Angeles.

Hu, J., Levy, R., & Schuster, S. (2022). Predicting scalar diversity with context-driven uncertainty over alternatives. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 68–74.

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

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

Lacina, R., Alexandropoulou, S., Ronai, E., & Gotzner, N. (2024). Scalar alternative activation in implicature processing: A lexical decision study with antonyms and negation. *PsyArXiv*. https://doi.org/10.31234/osf.io/r3q79

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and linguistic theory*, 587–610.

Leffel, T., Cremers, A., Gotzner, N., & Romoli, J. (2019). Vagueness in implicature: The case of modified adjectives. *Journal of Semantics*, *36*(2), 317–348.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, *176*, 1–14.

Ronai, E., & Xiang, M. (2022). Quantifying semantic and pragmatic effects on scalar diversity. *Proceedings of the Linguistic Society of America*, *7*(1), 5216.

Schmidt, L. A. (2009). *Meaning and compositionality as statistical induction of categories and constraints* [Doctoral dissertation, Massachusetts Institute of Technology].

Solt, S. (2009). Notes on the comparison class. *International workshop on vagueness in communication*, 189–206.

Solt, S. (2015). Measurement scales in natural language. *Language and Linguistics Compass*, *9*(1), 14–32.

Solt, S., & Gotzner, N. (2012). Experimenting with degree. *Semantics and linguistic theory*, 166–187.

van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of semantics*, *33*(1), 137–175.

Weicker, M., & Schulz, P. (2024). Children and adults privilege linguistic over visual information when creating comparison classes for prenominal gradable adjectives. *Glossa: a journal of general linguistics*, *9*(1).