Appendix

1 Multiple-statement presentation mode: Alexandropoulou & Gotzner's (accepted)

1.1 Alexandropoulou & Gotzner's (accepted) experimental material

Table A: Overview of relative adjectives in non-negated conditions (negated conditions included the same adjectives preceded by *not*) in Alexandropoulou & Gotzner's (accepted) Experiment 1. The top row presents the names of each adjective quadruple.

Item/ Condition	delicious	scalding	brilliant	sweltering	gorgeous	delighted	excellent	gigantic
Non-negated negative strong	*disgusting*	*freezing*	*idiotic*	*freezing*	*hideous*	*miserable*	*terrible*	*tiny*
Non-negated negative weak	*bland*	*cold*	*silly*	*cold*	*ugly*	*unhappy*	*bad*	*small*
Non-negated positive weak	*tasty*	*hot*	*intelligent*	*hot*	*pretty*	*happy*	*good*	*large*
Non-negated positive strong	*delicious*	*scalding*	*brilliant*	*sweltering*	*gorgeous*	*delighted*	*excellent*	*gigantic*

Table B: Overview of absolute adjectives in non-negated conditions (negated conditions included the same adjectives preceded by *not*) in Alexandropoulou & Gotzner's (accepted) Experiment 2. The top row presents the names of each adjective quadruple.

Item/ Condition	bolt upright	flawless	healthy	immaculate	pristine	safe	silky soft	spotless
Non-negated negative strong	*twisted*	*imperfect*	*sick*	*broken*	*filthy*	*dangerous*	*cracked*	*filthy*
Non-negated negative weak	*bent*	*impure*	*unwell*	*faulty*	*dirty*	*dodgy*	*rough*	*dirty*
Non-negated positive weak	*straight*	*pure*	*well*	*intact*	*clean*	*riskless*	*smooth*	*clean*
Non-negated positive strong	*bolt upright*	*flawless*	*healthy*	*immaculate*	*pristine*	*safe*	*silky soft*	*spotless*

One may observe that the quadruples **scalding** and **sweltering** for relative adjectives (Table A), and the quadruples **pristine** and **spotless** for absolute adjectives (Table B) have overlapping terms in all conditions except for the positive strong ones. However, we believe that the repeated items are not identical in all aspects. In the former case, while both quadruples share the single dimension of temperature, the repeated items do not match comprehensively. Their distinction lies in the varying standards established for each pair of overlapping terms: the standard value for the adjectives in the **sweltering** quadruple (i.e., contextually-defined degrees of weather temperature) is lower than the corresponding one determined by the modified noun in the **scalding** quadruple, which is water for cooking pasta. In the **spotless** quadruple, on the other hand, the repeated terms *clean, dirty, filthy* are assumed to be multidimensional (see, e.g., Sassoon 2012; Solt

2018). In contrast to the single dimension of temperature we discussed, what counts as *clean, dirty, filthy* is a function of various dimensions such as hygiene standards, sanitation and sterilization practices, the level of organization or orderliness in a space, overall maintenance of the physical space, etc. Crucially, different dimensions come into play when assessing the degree of cleanliness in the **pristine** and **spotless** items, with the evaluation centering on the cleanliness of different hospitals in the former and hotel rooms in the latter. Specifically, in the **pristine** quadruple, the cleanliness evaluation for each hospital is explicitly centered around its hygiene standards (cf. *decide which rating the hospital gets for its hygiene standards*, Table 2 of the paper), whereas in the **spotless** items this is made less specific (i.e., asking to decide which rating each applicant gets after cleaning a hotel room).

1.2 Adjective selection procedure

The following pairs of informationally weak and strong adjectives were adapted from Gotzner et al.'s (2018) experimental materials:

- <*tasty, delicious*>
- <*cold, freezing*>
- <*hot, scalding*>
- <*idiotic, silly*>
- <*intelligent, brilliant*>
- <*ugly, hideous*>
- <*pretty, gorgeous*>
- <*unhappy, miserable*>
- <*happy, delighted*>
- <*good, excellent*>
- <*small, tiny*>
- <*large, gigantic*>
- <*bent, twisted*>
- <*dirty, filthy*>
- <*clean, spotless*>

The remaining pairs of informationally weak and strong adjectives tested in Alexandropoulou & Gotzner's (accepted) experiments (as well as in Experiments 1 and 2 of the present paper) are the following:
<*bland, disgusting*>, <*hot, sweltering*>, <*bad, terrible*>, <*straight, bolt upright*>, <*imperfect, impure*>, <*pure, flawless*>, <*unwell, sick*>, <*well, healthy*>, <*faulty, broken*>, <*intact, immaculate*>, <*clean, pristine*>, <*dodgy, dangerous*>, <*riskless, safe*>, <*rough, cracked*>, <*smooth, silky soft*>.

The above adjective pairs were chosen on the basis of the relevant lemmas from online dictionaries of English that often provide or imply a strength difference between weak and strong expressions (e.g., *terrible*: 'extremely bad', *pristine*: 'clean and fresh as if new, spotless'). Dictionary lemmas further include antonyms of weak (e.g., *well* vs. *unwell*) and strong adjectives (e.g., *healthy* vs. *sick*). The selected antonymic pairs were moreover checked against a list of common opposites available on the web.[1] Lastly, the selected adjectives and groupings thereof were double-checked by native speakers of English.

[1] https://www.enchantedlearning.com/wordlist/opposites.shtml

In particular, for the weak and strong adjectives we grouped together (Horn entailment scales), we applied the following diagnostic: *x is* ADJ_{WEAK} *but not* ADJ_{STRONG} should be OK, whereas *x is* ADJ_{STRONG} *but not* ADJ_{WEAK} should not be OK, e.g., *rough but not cracked* vs. *#cracked but not rough*.

For categorizing the adjectives as being relative vs. absolute gradable adjectives, we consulted Gotzner et al.'s (2018) experimental materials and Stephanie Solt's own adjective dataset (p.c.), we also borrowed examples of different gradable adjective classes from the literature, and further applied the diagnostic that follows: the licit use of the adverbs *very* and *extremely* with a gradable adjective should be an indication of a relative, as opposed to absolute, adjective.

As far as evaluative polarity is concerned, English native speakers' informal judgements confirmed the selection of the pairs of antonymic adjectives in the given contexts and in combination with the specific nouns. The characterization of an adjective in terms of evaluative polarity was further in agreement with the valence scores in Mohammad's (2018) NRC Valence, Arousal, and Dominance (VAD) Lexicon.

As becomes evident, in a few instances, we used morphological antonyms—either cases adapted from Gotzner et al.'s (2018) material that are widely discussed in the relevant literature (e.g., *happy/unhappy* in Horn 1989; Krifka 2007; Tessler & Franke 2018), or recommended by English native speakers' informal judgements and dictionaries (*well/unwell, impure, imperfect, bolt upright, silky soft*)—as the sole or optimal antonyms. However, we do not think that the observed interpretation patterns are to be attributed to the use of morphological antonyms. It is worth mentioning that Ruytenbeek et al. (2017) found a more pronounced polarity asymmetry with morphological antonyms, but they still identified its availability with non-morphological antonyms consisting of relative adjectives. Regarding periphrastic items (*bolt upright, silky soft*), the relevant interpretation patterns, both with and without negation (see Figure F), closely resembled those of non-periphrastic items, suggesting that their inclusion had no noticeable impact.

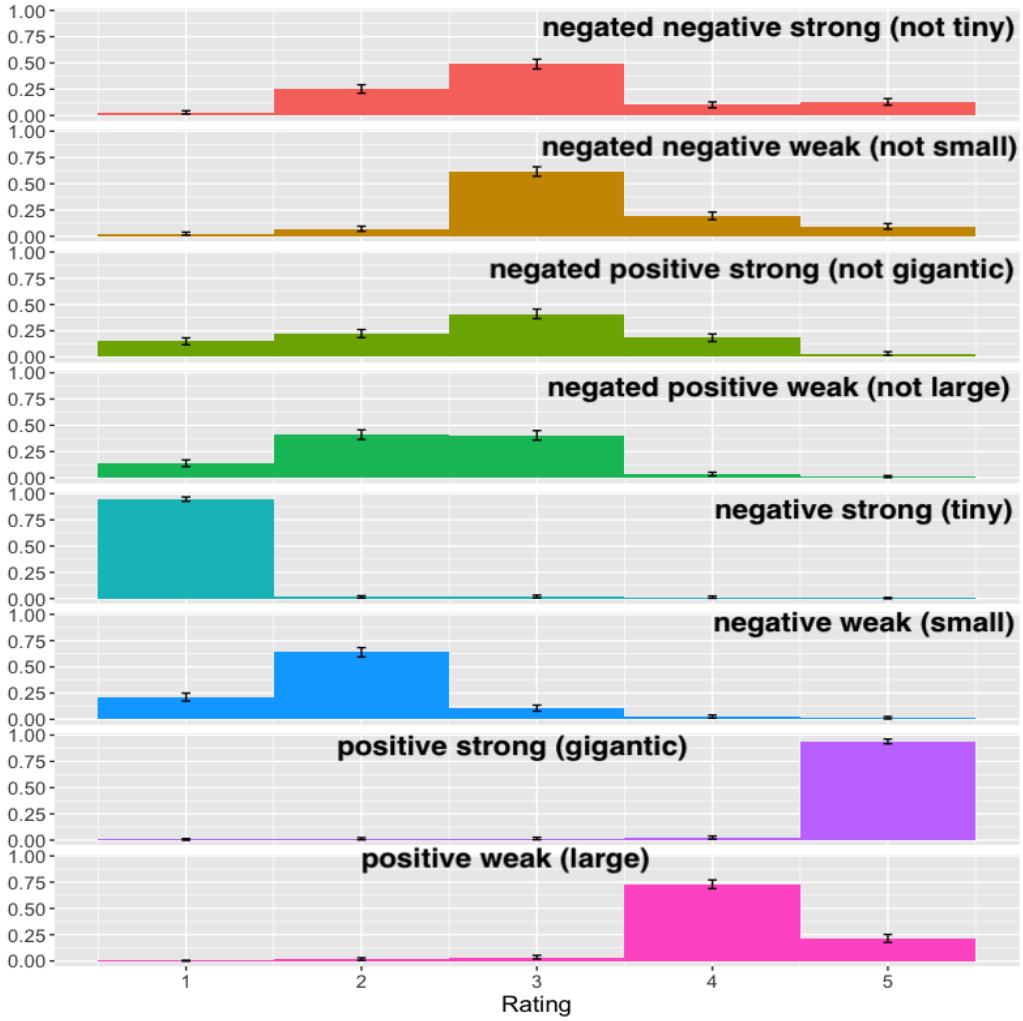
1.3 Alexandropoulou & Gotzner's (accepted) design and results

Alexandropoulou & Gotzner (accepted) manipulated the factors Informational/Scalar Strength (informationally weak vs. strong), Evaluative Polarity (evaluatively positive vs. negative), and Negation (presence vs. absence of negation).

They ran cumulative link mixed-effects models for relative and absolute adjectives separately (see relevant data in Figures A and B, respectively). In both models, they included the treatment-coded fixed effects of Negation (with negated as the reference level), and Scalar Strength (with weak as the reference level), and the sum-coded fixed effect of Evaluative polarity. They further included the maximal converging random-effect structure justified by their experimental design, with random by-participant and by-item intercepts and slopes. Tables C and D summarize the output of the relevant models for relative and absolute gradable adjectives, respectively.

The statistically significant simple effect of Negation revealed by the analysis of relative adjectives indicates the availability of an asymmetric interpretation pattern for negated positive (*not large*) and negative weak relative adjectives (*not small*) as compared to their base forms without negation (*large* and *small*, respectively). The statistically significant interaction of Negation*ScalarStrength indicates that an asymmetric interpretation pattern is less likely to arise for positive and negative strong relative terms (*not gigantic* vs. *not tiny*) than for weak relative terms (*not large* vs. *not small*). The analysis of absolute adjectives, on the other hand, revealed no significant effect of Negation or interaction

Figure A: Proportions per rating per adjective condition in Alexandropoulou & Gotzner's (accepted) experiment on relative adjectives (multiple-statement presentation mode). Error bars represent 95% Confidence Intervals.



of Negation*ScalarStrength. Hence, Alexandropoulou & Gotzner (accepted) found no evidence of an asymmetric interpretation of positive (*not clean*) and negative weak absolute adjectives (*not dirty*) in the scope of negation, nor that this may be different for positive and negative strong terms (*not pristine* vs. *not filthy*). These results are in line with the evident symmetric response patterns of negated positive and negative weak absolute conditions in Figure B, and the largely overlapping response patterns of negated weak and strong absolute conditions.

Figure B: Proportions per rating per adjective condition in Alexandropoulou & Gotzner's (accepted) experiment on absolute adjectives (multiple-statement presentation mode). Error bars represent 95% Confidence Intervals.

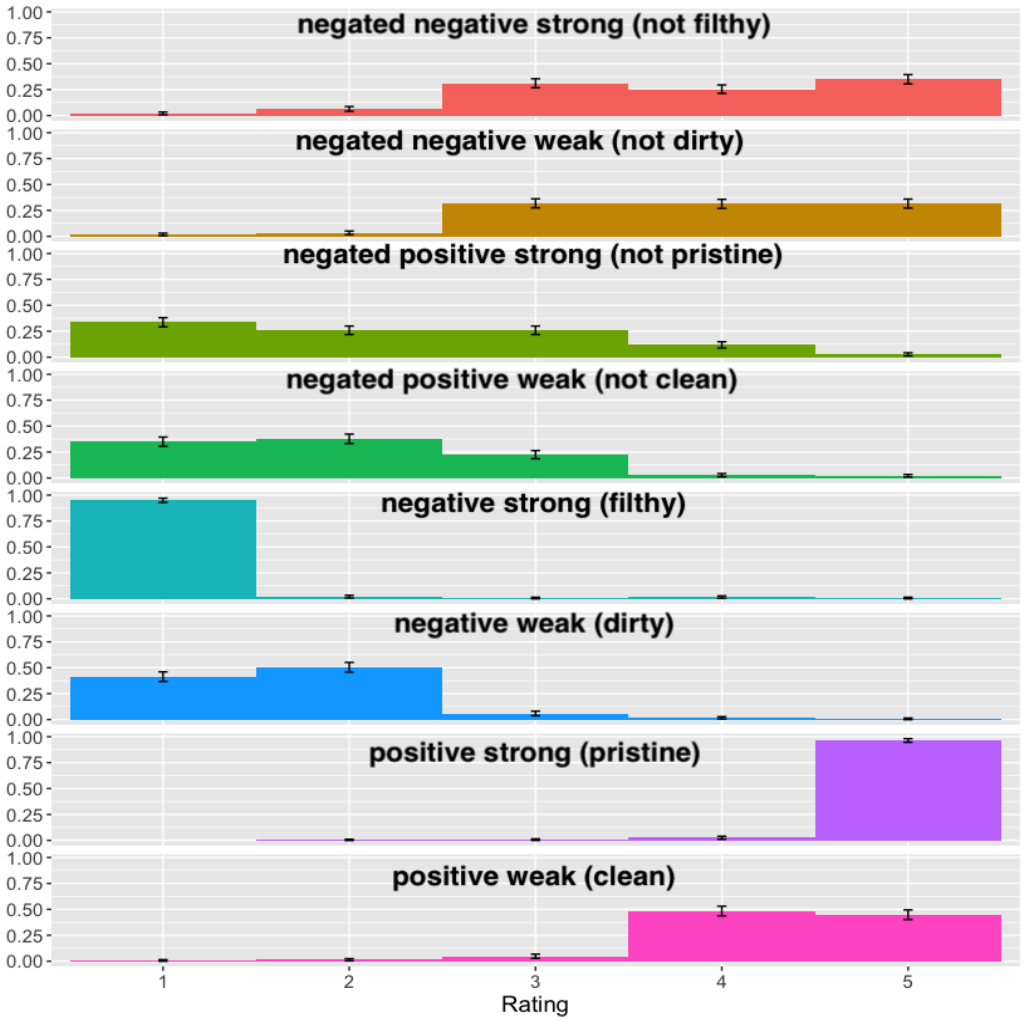


Table C: Alexandropoulou & Gotzner's (accepted) multiple-statement presentation mode experiment on relative adjectives: Output of cumulative link model.
clmm(Rating ~ Negation * ScalarStrength * Polarity + (Negation + ScalarStrength * Polarity | Participant) + (Negation * ScalarStrength * Polarity | Item), data =data)

	Estimate	SE	z-value	p-value
NegationNon-negated	0.64392	0.21714	2.965	0.00302 **
ScalarStrengthStrong	0.16399	0.09742	1.683	0.09231 .
Polarity1	-1.17430	0.13283	-8.841	< 2e-16 ***
NegationNon-negated:ScalarStrengthStrong	-0.63507	0.30563	-2.078	0.03772 *
NegationNon-negated:Polarity1	3.96282	0.17085	23.194	< 2e-16 ***
ScalarStrengthStrong:Polarity1	0.56976	0.23580	2.416	0.01568 *
NegationNon-negated:ScalarStrengthStrong:Polarity1	4.56495	0.28776	15.864	< 2e-16 ***

Table D: Alexandropoulou & Gotzner's (accepted) multiple-statement presentation mode experiment on absolute adjectives: Output of cumulative link model.
clmm(Rating ~ Negation * ScalarStrength * Polarity + (Negation + ScalarStrength * Polarity | Participant) + (Negation * ScalarStrength * Polarity | Item), data =data)

	Estimate	SE	z-value	p-value
Negationnon-negated	0.17288	0.12453	1.388	0.165
ScalarStrengthStrong	0.20254	0.15342	1.320	0.187
Polarity1	-2.06506	0.16314	-12.658	< 2e-16 ***
Negationnon-negated:ScalarStrengthStrong	0.36063	0.31877	1.131	0.258
Negationnon-negated:Polarity1	4.92392	0.22361	22.020	< 2e-16 ***
ScalarStrengthStrong:Polarity1	0.02906	0.32270	0.090	0.928
Negationnon-negated:ScalarStrengthStrong:Polarity1	4.59050	0.27229	16.859	< 2e-16 ***

2 Single-statement presentation mode experiments

The sample size, the participant exclusion criterion, the experimental procedure, the hypotheses, and analyses of Experiments 1 and 2 were pre-registered and are available on the Open Science Framework.[2] Due to a reviewer's suggestion to Alexandropoulou & Gotzner (accepted), we report a statistical analysis that is different from the pre-registered one and that captures the critical contrasts we are after in a simpler model.

2.1 Design

The analyses of Experiment 1 on relative adjectives and Experiment 2 on absolute adjectives were the same, including the same 3 manipulations: (i) Informational/Scalar Strength (informationally weak vs. strong), (ii) Evaluative Polarity (evaluatively positive vs. negative), and (iii) Negation (non-negated vs. negated).

2.2 Procedure

Experiment 1 demo link: https://farm.pcibex.net/r/rJAcWR/experiment.html?test=true
Experiment 2 demo link: https://farm.pcibex.net/r/lBwzet/experiment.html?test=true

The antonymic adjective pairs used in Experiment 1's practice trials were: *distasteful* vs. *exquisite*, *awful* vs. *marvelous*.

The antonymic adjective pairs used in Experiment 2's practice trials were: *erroneous* vs. *correct*, *wrong* vs. *right*.

[2] Pre-registrations to be found on the following links: Experiment 1: https://osf.io/59bwd/?view_only=3544fddd3af24b5ab39fe08e0c8803d1, Experiment 2: https://osf.io/crf2t/?view_only=9c3e9af69c3c4dbaa79519d2897b0b25.

Example practice item from Experiment 1 (relative adjectives)

Context:
Martin is learning how to bake. His mother gives him feedback.
Please decide which rating each of Martin's cakes gets in relation to its taste based on his mother's feedback.
1 = distasteful; 5 = exquisite

Martin's mother says:

The chocolate cake was exquisite.

1 2 3 4 5

The cheesecake was distasteful.

1 2 3 4 5

Example practice item from Experiment 2 (absolute adjectives)

Context:
Yesterday, the class had its final math test of the year. The teacher gives the students feedback.
Please decide which rating each student gets for their math test based on the teacher's feedback.
1 = erroneous; 5 = correct

The teacher says:

Peter's calculation was erroneous.

1 2 3 4 5

Anna's calculation was correct.

1 2 3 4 5

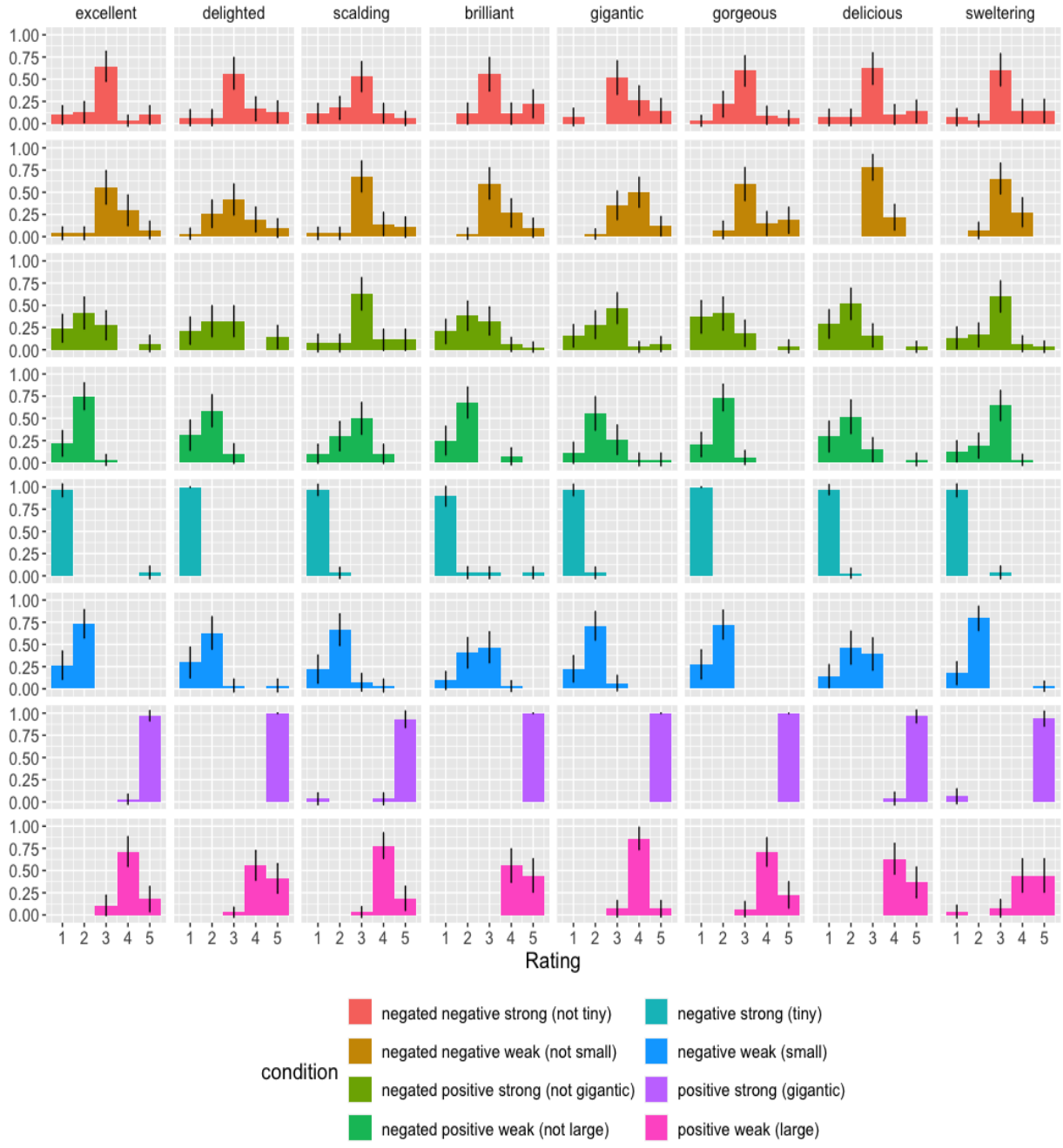
2.3 Results

For both Experiment 1 and 2, we ran the same cumulative link mixed-effects models as those in Alexandropoulou & Gotzner accepted.

2.3.1 Experiment 1: Relative adjectives

Figure D presents the proportions per rating per relative adjective condition for individual contexts/items in single-statement presentation mode.

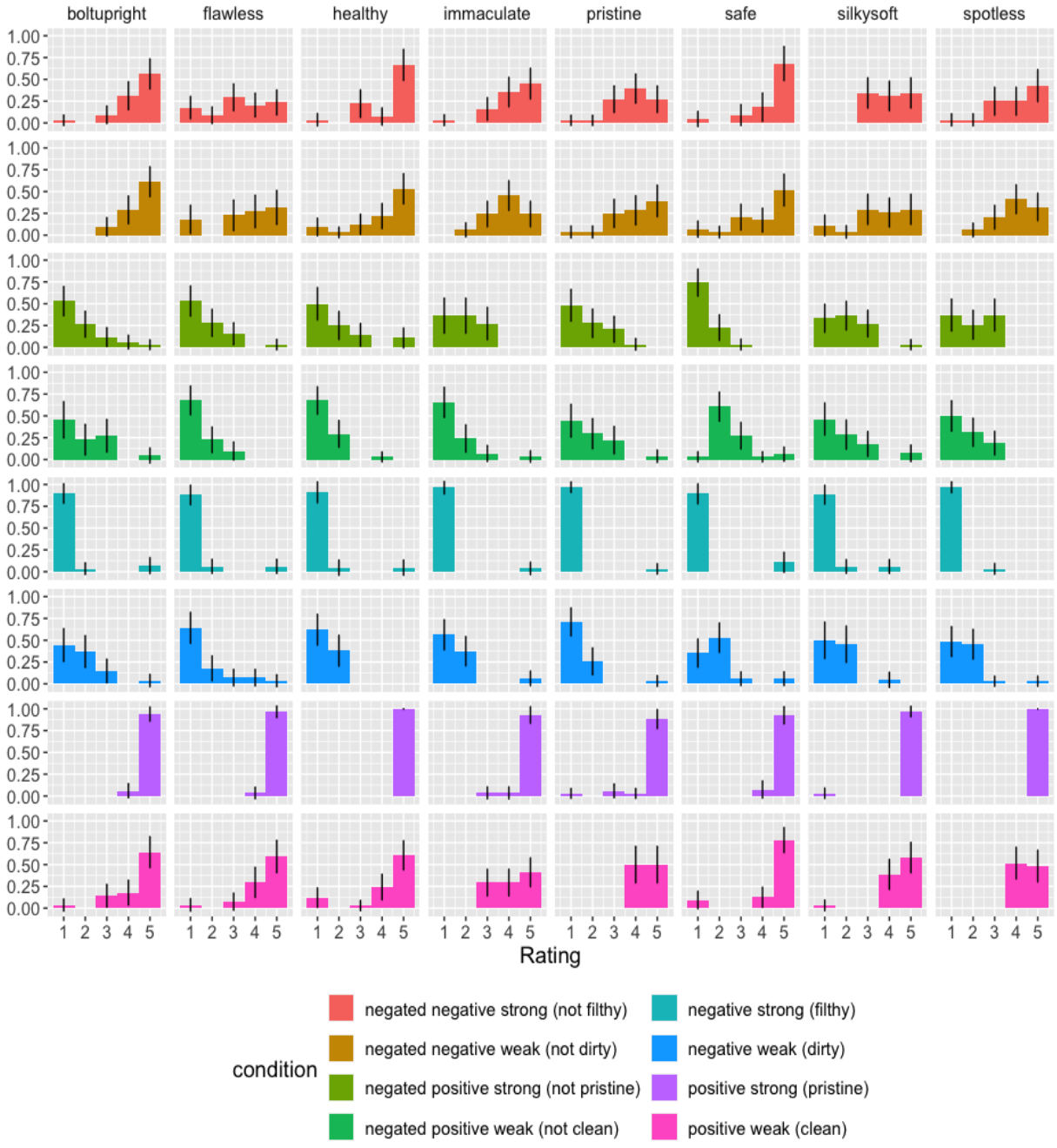
Figure D: Proportions per rating per adjective condition for individual contexts/items in Experiment 1 (relative adjectives) with single-statement presentation mode. Error bars represent 95% Confidence Intervals. Contexts are named after the respective positive strong terms.



2.3.2 Experiment 2: Absolute adjectives

Figure F presents the proportions per rating per absolute adjective condition for individual contexts/items in single-statement presentation mode.

Figure F: Proportions per rating per adjective condition for individual contexts/items in Experiment 2 (absolute adjectives) with single-statement presentation mode. Error bars represent 95% Confidence Intervals. Contexts are named after the respective positive strong terms.



References

Alexandropoulou, Stavroula & Gotzner, Nicole. accepted. The interpretation of relative and absolute adjectives under negation. *Journal of Semantics*.

Gotzner, Nicole & Solt, Stephanie & Benz, Anton. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659. https://doi.org/10.3389%2Ffpsyg.2018.01659.

Horn, Laurence R. 1989. *A natural history of negation*. Chicago: University of Chicago Press.

Krifka, Manfred. 2007. Negated antonyms: Creating and filling the gap. In Sauerland, Uli & Stateva, Penka (eds.), *Presupposition and implicature in compositional semantics*, 163–177. Houndmill: Palgrave Macmillan.

Ruytenbeek, Nicolas & Verheyen, Steven & Spector, Benjamin. 2017. Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics* 2(1). 1–27. https://doi.org/10.5334/gjgl.151.

Sassoon, Galit W. 2012. A typology of multidimensional adjectives. *Journal of Semantics* 30(3). 335–380. https://doi.org/10.1093/jos/ffs012

Solt, Stephanie. 2018. Multidimensionality, subjectivity and scales: Experimental evidence. In Castroviejo, Elena & McNally, Louise & Sassoon, Galit Weidman (eds.), *The Semantics of Gradability, Vagueness, and Scale Structure*, 55–91. Springer International Publishing AG.

Tessler, Michael H. & Franke, Michael. 2018. Not unreasonable: Carving vague dimensions with contraries and contradictions. In Kalish, Charles & Rau, Martina & Zhu, Jerry & Rogers, Timothy (eds.), *Proceedings of CogSci*, vol. 40. 1108–1113.