Exclusives vary in strength and scale structure: experimental evidence *

Eszter Ronai[1] and Lucas Fagen[2]

[1] Northwestern University, Evanston, Illinois, U.S.A.
ronai@northwestern.edu
[2] The University of Chicago, Chicago, Illinois, U.S.A.
lfagen@uchicago.edu

Abstract

This paper is an investigation of parameters of variation across English exclusive modifiers. We report on three experiments that test the robustness of exclusionary inference calculation (e.g., *merely intelligent* → *not brilliant*) across a large number of different lexical scales. Our findings reveal that 1) *just* excludes less robustly than *only*, 2) while *only* allows both complement-exclusion and rank-order readings, *merely* prefers rank-order ones, and 3) *just* and *only* are equally QUD-sensitive. We discuss these results in light of existing theoretical proposals about the semantics of exclusives.

1 Background

1.1 Exclusive modifiers

Exclusive modifiers, which in English include *only*, *just*, and *merely* (1), form a lexical class (Coppock and Beaver 2014), conveying that some proposition is true (the *prejacent*, (1-a)) and that alternatives to the prejacent are false (1-b).

(1) Mary **only/just/merely** ate the cookies.
   a. → Mary ate the cookies
   b. → Mary ate nothing other than the cookies

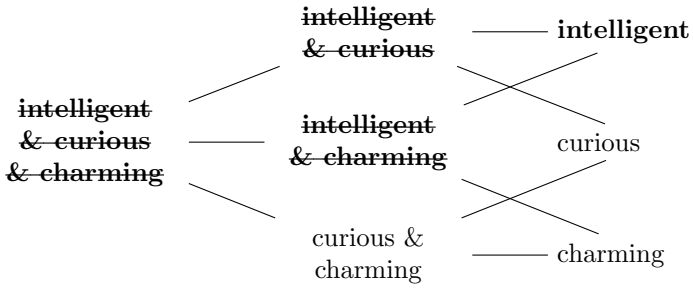
(2) The student is **only** intelligent.
   a. → The student is not curious, not charming, etc.
   b. → The student is not brilliant

Exclusives vary along different parameters. In this paper we focus on two: scale structure bias and strength of exclusion. It has been proposed that the excluded alternative set can vary in scale structure (Klinedinst 2005; Beaver and Clark 2008; Coppock and Beaver 2014). Specifically, exclusives can have 'complement-exclusion' readings (2-a) which exclude everything other than the prejacent, and 'rank-order' readings (2-b) which exclude alternatives ranked higher than the prejacent on a scale. These readings can receive a uniform semantic analysis, e.g. Coppock and Beaver (2014) analyze complement-exclusion and rank-order readings as excluding along differently structured scales (as shown in (3) and (4), adapted from their examples 27 and 29). But what we are primarily interested in in this paper is that exclusives do not always vary freely along this parameter: while *only* and *just* admit both readings, *merely* has been argued to prefer rank-order scales. Coppock and Beaver (2014) analyze variation in scale structure as resulting from "soft preferences" rather than absolute restrictions. Such effects are likely to

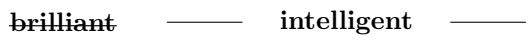
*We are grateful to three anonymous AC reviewers and the audience at the UChicago LEAP Workshop for their feedback. This material is based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. Authors contributed equally to this work and are listed in reverse alphabetical order.

emerge more starkly across items in an experimental setting than via intuition alone, and are therefore worth testing directly.

(3)



(4)



The strength of the exclusion has also been argued to vary. *Just* but not *only* has a wider range of readings, paraphrasable with *simply*, that exclude alternatives understood as uninformative, unknown (5-a), redundant (5-b), or irrelevant, but not necessarily false. To explain this, various authors have proposed that *just* can operate on alternatives that *only* does not, including: covert causal modifiers with trivial semantic content (Wiegand 2018), answers to 'potential' questions in addition to the current QUD (Warstadt 2020), or metalinguistic alternatives at the speech act level (Beltrama 2022).

(5) a. The lights in this place **just**/**#only** turn off and on. (Warstadt 2020, ex. 1a)
Paraphrase: the lights turn off and on for no reason.

b. The pumpkin bisque is **just**/**#only** delicious! (Warstadt 2020, ex. 1b)
Paraphrase: the pumpkin bisque is extremely delicious.

Warstadt (2020) argues that such readings also require relaxing the truth-conditional status of the exclusive operation, proposing a distinction between "strong" exclusives which declare alternative propositions false, and "weak" exclusives which declare them unassertable. (See also Beltrama 2022, who proposes that emphatic *just* as in (5-b) declares alternatives not unassertable per se but not "worthy of assertion", p. 347.) Warstadt (2020) analyzes *only* as strong and *just* as weak. This is contra Coppock and Beaver (2014), who take both *only* and *just* to be strong, and have no explanation for weak readings with *just*. Warstadt's theoretical proposal regarding strength of exclusion is therefore novel and also worth testing experimentally.

1.2 Scalar diversity

Our testing ground in this paper is pairs of lexical items that form a scale. More specifically, we turn to the scalar diversity phenomenon: the observation that scalar expressions vary in how likely they are to lead to scalar implicature (SI) (i.a. van Tiel et al. 2016). A classic example of SI is (6): upon encountering an utterance containing *some*, hearers compute the negation of its stronger scalar alternative *all*. Similarly to ⟨*some*, *all*⟩, e.g. ⟨*intelligent*, *brilliant*⟩ also forms a scale: an utterance containing *intelligent* can lead to the SI *not brilliant* (7). But variation exists across different scales: the SI in (6) arises much more robustly than the one in (7).

(6) Mary ate some of the cookies.
→ SI: Mary ate some, but not all, of the cookies.

(7) The student is intelligent.
→ SI: The student is intelligent, but not brilliant.

Scalar diversity persists even in the presence of exclusives. Ronai and Xiang (2022) (henceforth R&X) found that even when sentences such as (6)-(7) contain *only*, variation still remains

in the likelihood of deriving a *not all* or *not brilliant* inference. This is puzzling, since while SI is a cancellable pragmatic inference, *only* encodes alternative exclusion in the semantics—which would predict uniformly ceiling-level inference rates. R&X hypothesized that interpretations of *only* were split between rank-order and complement-exclusion readings, leading to variation in whether the stronger scalar term was included in the alternative set. Given *The student is only intelligent*, the *not brilliant* inference would arise with rank-order *only*, but not necessarily with complement-exclusion *only*, which could be understood as excluding other unrelated properties (*not curious*, *not charming*, etc).

2 Experiments 1 and 2: *just* and *merely*

Though recent progress has been made in the theoretical literature describing variation among exclusives, much remains to be understood about which parameters vary and how. Here, we provide the first systematic experimental assessment of this domain, focusing on strength of exclusion and scale structure bias. First, we test Warstadt (2020)'s claim about strength of exclusion by comparing *just* vs. *only*. Second, we compare *only* vs. *merely*, the latter of which is claimed to prefer rank-order readings. This also addresses R&X's hypothesis that interpretations of *only* were split according to scale structure in their experiment.

2.1 Task and procedure

Experiment 1 was conducted on the web-based PCIbex platform (Zehr and Schwarz 2018). 40 speakers of American English were recruited on Prolific and were screened with a demographic survey and attention checks; data from 39 participants is reported below. The experiment used an inference task (i.a. van Tiel et al. 2016): participants saw sentences such as "Mary: *The student is just intelligent.*" and were asked the question "Would you conclude from this that Mary thinks the student is not brilliant?". They responded with "Yes", which indexes that the participant has calculated the exclusionary (*not brilliant*) inference, or "No", which suggests that the inference was not calculated. We used the same task and items as R&X's experiments, allowing for a direct comparison to their results. We tested 51 different lexical scales.

Experiment 2 was identical to Experiment 1 in its basic procedure, task and items, as well as participant recruitment and removal. Data from 35 participants is reported. Experiment 2 tested the exclusive *merely*. That is, participants saw stimuli such as "Mary: *The student is merely intelligent*".

2.2 Predictions

We make the following two predictions. First, given Warstadt (2020)'s claim that *just* is a weak exclusive, while *only* is a strong exclusive, we predict lower rates of inference calculation for Experiment 1 than was found for *only* (by R&X). Second, we predict higher rates of inference calculation for Experiment 2 than was found for *only*. This is because all our items test rank-order alternatives, and while *only* allows both complement-exclusion and rank-order readings, *merely* has been claimed to prefer rank-order readings (Coppock and Beaver 2014).

2.3 Results

For reasons of space, we present the visualizations of our experimental results in the Appendix. Figure 1 shows the results of both experiments, along with R&X's SI and *only* results. The

percent of exclusionary inference calculation corresponds to the percent of "Yes" responses in the inference task. To compare the rates of inference calculation in Experiments 1-2 to R&X's *only* experiment, we fit a logistic mixed effects regression model using the lme4 package in R (Bates et al. 2015). The model predicted Response ("Yes" vs. "No") as a function of Exclusive (*just* vs. *only* vs. *merely*). Random intercepts for participants and random slopes and intercepts for items were included. Since our predictions concern comparing *just* and *merely* to *only*, the predictor Exclusive was treatment coded, with *only* coded as the reference level. The model revealed significantly lower rates of inference calculation with *just* compared to *only* (Estimate=-0.7, SE=0.28, z=-2.5, p <0.05), as well as significantly higher rates of inference calculation with *merely* than was found with *only* (Estimate=0.96, SE=0.28, z=3.38, p <0.001). Averaged over the 51 different scales, the target exclusionary inference was calculated at the rate of 52.9% with *just*, 65.5% with *only*, and 80.2% with *merely*.

Since inference rates were lowest with *just*, the question may arise whether it can be maintained that *just* excludes alternatives semantically. To test this, we fit an additional statistical model comparing our Experiment 1 findings to R&X's SI experiment, which found an average rate of 33.1% SI calculation. The fixed effects predictor was sum-coded (SI: -0.5 and *just*: 0.5). SI rates were found to be significantly lower than the current Experiment 1 results with *just* (Estimate=1.32, SE=0.25, z=5.35, p <0.001). This confirms that alternative exclusion with all three exclusive modifiers is stronger than alternative exclusion via SI.

2.4 Discussion

Both predictions we made for Experiments 1-2 are borne out by the results. First, Experiment 1 found lower rates of exclusionary inference calculation with *just* than with *only*. This is consistent with the hypothesis that *just* excludes alternatives via a weaker semantic operation than *only* —a question we return to in Experiment 3. Second, Experiment 2 found higher inference rates with *merely* than with *only*. Since all our items test rank-order alternatives, this strongly supports the claim that *merely* biases toward rank-order scales (Coppock and Beaver 2014). These results also support R&X's hypothesis that participants in their experiment were split between rank-order and complement-exclusion readings of *only*. When a rank-order bias was introduced by *merely*, the stronger scalar term was more unambiguously understood as one of the salient alternatives, which led to an increase in calculating the target inference.

One may wonder whether different scales interact differently with the two tested parameters of variation. In order to check whether the relative order of different lexical scales remained consistent across manipulations, we calculated rank-order correlations using Kendall's τ_B, and found significant by-item correlations between experiments. Items with low SI rates continue to have relatively low inference rates even with (stronger) exclusives. As SI rates increase, so do inference rates with *just* (τ_B=0.59, p<0.001); as rates with *just* increase, so do rates with *only* (τ_B=0.59, p<0.001); and rates with *merely* are also correlated with *only* (τ_B=0.53, p<0.001). Only a small minority (≈5) of scales deviate from the general patterns. This highlights the role of lexico-semantic factors in the scalar diversity phenomenon.

3 Experiment 3: *just* + QUD

One explanation for our Experiment 1 results could be that *just* excludes alternatives via a weaker semantic operation than *only* (as suggested by Warstadt 2020, Beltrama 2022). If *just* were declaring scalar alternatives like *brilliant* uninformative or unassertable, rather than false, participants might have been more reluctant to answer "Yes" to questions like "Would you

conclude from this that Mary thinks the student is not brilliant?". Another explanation could be that *just* excludes a wider range of nonfocal alternatives. Consider Warstadt's proposal that *just* can answer potential questions ("intuitively possible future QUDs", p. 373) in addition to the current QUD. According to Warstadt, *just* in (8-a) signals that there are no assertable answers to the potential followup (8-b), "preventing the addressee from asking a useless question" (p. 373). On this account, *brilliant* or any other alternative to *intelligent* would not have been one of the excluded alternatives in our Experiment 1, because *just* was answering a potential question rather than the QUD itself.

(8) a. The lights in this place just turn off and on.
b. Why do the lights turn off and on?

A third possibility is that theories like Wiegand (2018)'s and Warstadt (2020)'s that aim to unify canonically exclusive and noncanonical readings of *just* under a single entry are on the wrong track. Instead, the lower rates of target inference calculation in Experiment 1 could reflect lexical ambiguity: participants excluded stronger scalar alternatives like *brilliant* when *just* was interpreted exclusively, and not otherwise.

To test these possibilities, we turn to another finding from the experimental pragmatics literature: explicit QUDs (Roberts 1996/2012) encourage inference calculation (i.a. Degen 2013; Zondervan, Meroni, and Gualmini 2008; Ronai and Xiang 2021; Ronai and Xiang 2022). For example, R&X found that when sentences with *only* were presented as answers to polar questions containing the stronger scalar term, participants were more likely to calculate the target exclusionary inference than in the null context *only* experiment. Given Warstadt's proposal that *just* can answer potential questions other than the current QUD, we would predict a differential effect of experimentally manipulating the QUD. Specifically, we would predict *just* to exhibit reduced sensitivity to the QUD compared to *only*.

3.1 Task and procedure

Experiment 3 had the same basic procedure, task, items, and participant recruitment and exclusion as the previous experiments. Data from 39 participants is reported. Similarly to Experiment 1, we tested the exclusive *just* in an inference task. But in this experiment, we embedded Mary's statements in a dialogue context, where another conversational participant, Sue, first asked a question. Sue's questions were polar questions that contained the stronger scalar term, while Mary's answers were slightly modified from Experiment 1 to ensure dialogue coherence (e.g., *the student* was changed to *he*). An example is shown below:

(9) Sue: Is the student brilliant?
Mary: He is just intelligent.

Participants again answered task questions like "Would you conclude from this that Mary thinks the student is not brilliant?" with either "Yes" or "No".

3.2 Predictions

To assess whether exclusives differ from each other in how sensitive they are to the QUD, we will compare the findings in Experiment 3 (*just* + QUD) to Experiment 1 (*just*) and R&X's *only* and *only* + QUD experiments. In their *only* + QUD experiment, experimental items were identical to (9), except Mary's answers contained *only*. We make the following predictions for the comparison of the four experiments. First, as we saw in Section 2, rates of calculating the target exclusionary inference should be higher with *only* than *just*. Second, since QUDs

generally encourage inference calculation, we predict higher rates for QUD experiments than for null context experiments. Lastly and most crucially, given Warstadt's proposal that *just* can answer non-QUD potential questions, we predict an interaction of exclusives and context, such that adding the QUD has less of an effect on *just* than on *only*.

3.3 Results

Figure 2 shows the results of Experiment 3 (*just* + QUD), along with Experiment 1 (*just*, repeated from Figure 1), as well as R&X's *only* (repeated from Figure 1) and *only* + QUD results. Averaged over the different scales, the target exclusionary inference was calculated at the rate of 78.7% in Experiment 3 —compare R&X's *only* + QUD experiment, which had a rate of 88.3%, as well as the null context exclusives from Section 2.3.

For the statistical analysis of the four experiments, we fit a logistic mixed effects regression model that predicted Response ("Yes" vs. "No") as a function of Context (QUD vs. null context), Exclusive (*just* vs. *only*) and their interaction. The maximal converging random effects structure (Barr et al. 2013) included random intercepts for participants and items, as well as random slopes for "Context" and "Exclusive" (but not their interaction) for items. The fixed effects predictors we sum-coded before analysis (null context: -0.5 and QUD: 0.5; *just*: -0.5 and *only*: 0.5). The analysis revealed a significant effect of Context, such that QUD experiments led to higher rates of calculating the target inference than null context experiments (Estimate=1.84, SE=0.25, z=7.39, p <0.001) and a significant effect of Exclusive, such that *only* experiments led to higher rates than *just* experiments (Estimate=0.86, SE=0.25, z=3.47, p <0.001). However, the interaction was not significant (Estimate=0.18, SE=0.46, z=0.39, p =0.7).

3.4 Discussion

We did not find the statistical interaction predicted from Warstadt's proposal, namely that *just* would be less sensitive to the QUD manipulation than *only*. We interpret this as speaking against a theory that proposes one unified semantics for exclusive and nonexclusive *just*, where this unified *just* can answer potential questions that are not the current QUD. Instead, we tentatively propose that our results are most compatible with there being several different lexical entries for *just*. It is possible that exclusive *just* answers the QUD, while other flavors of *just* are not exclusive and do not answer the QUD. Adding the explicit QUD could have increased the rate of QUD-sensitive *just* interpretations compared to Experiment 1, because participants assumed that Sue's question was relevant. This in turn led to a higher rate of exclusionary inference calculation, since the only flavor of *just* that answers the QUD is exclusive *just*. In this way a lexical ambiguity account would predict our finding that the QUD manipulation did raise inference rates for *just*, but that we did not find *just* to be less QUD-sensitive than *only*.

4 Conclusion

We presented novel experimental evidence testing variation across exclusive modifiers in English. Our results provided support for two claims from the theoretical literature: that exclusives vary in scale structure bias (with *merely* preferring rank-order readings) and in strength of exclusion (with *just* excluding less robustly). However, we did not find evidence for exclusives varying in their QUD-sensitivity, and in particular that *just* can answer questions that are not the current QUD. Our QUD results tentatively suggest that the apparent strength difference between *just* and *only* does not stem from variation in logical strength but instead from lexical ambiguity.

References

Barr, Dale J et al. (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". In: *Journal of Memory and Language* 68.3, pp. 255–278. DOI: 10.1016/j.jml.2012.11.001.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Beaver, David and Brady Clark (2008). *Sense and sensitivity: how focus determines meaning*. Chichester: Wiley-Blackwell.

Beltrama, Andrea (2022). "Just perfect, simply the best: an analysis of emphatic exclusion". In: *Linguistics and Philosophy* 45, pp. 321–364. DOI: 10.1007/s10988-021-09326-x.

Coppock, Elizabeth and David Beaver (2014). "Principles of the exclusive muddle". In: *Journal of Semantics* 31.3, pp. 371–432. DOI: 10.1093/jos/fft007.

Degen, Judith (2013). "Alternatives in Pragmatic Reasoning". PhD thesis. University of Rochester.

Klinedinst, Nathan (2005). "Scales and *only*". MA thesis. UCLA.

Roberts, Craige (1996/2012). "Information structure in discourse: Towards an integrated formal theory of pragmatics". In: *Semantics and Pragmatics* 5.6, pp. 1–69. DOI: 10.3765/sp.5.6.

Ronai, Eszter and Ming Xiang (2021). "Pragmatic inferences are QUD-sensitive: an experimental study". In: *Journal of Linguistics* 57.4, pp. 841–870. DOI: 10.1017/S0022226720000389.

— (2022). "Quantifying semantic and pragmatic effects on scalar diversity". In: *Proceedings of the Linguistic Society of America*. Vol. 7. 1, p. 5216. DOI: 10.3765/plsa.v7i1.5216.

van Tiel, Bob et al. (2016). "Scalar diversity". In: *Journal of Semantics* 33.1, pp. 137–175.

Warstadt, Alex (2020). ""Just" don't ask: exclusives and potential questions". In: *Proceedings of Sinn und Bedeutung 24*. Ed. by Michael Franke et al. Vol. 2, pp. 373–390.

Wiegand, Mia (2018). "Exclusive morphosemantics: *just* and covert quantification". In: *Proceedings of the 35th West Coast Conference on Formal Linguistics*. Ed. by Wm. G. Bennett, Lindsay Hracs, and Dennis Ryan Storoshenko. Somerville, MA, pp. 419–429.

Zehr, Jeremy and Florian Schwarz (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832.

Zondervan, Arjen, Luisa Meroni, and Andrea Gualmini (2008). "Experiments on the Role of the Question Under Discussion for Ambiguity Resolution and Implicature Computation in Adults". In: *Proceedings of Semantics and Linguistic Theory (SALT) 18*. Ed. by Tova Friedman and Satoshi Ito, pp. 765–777.

Appendix

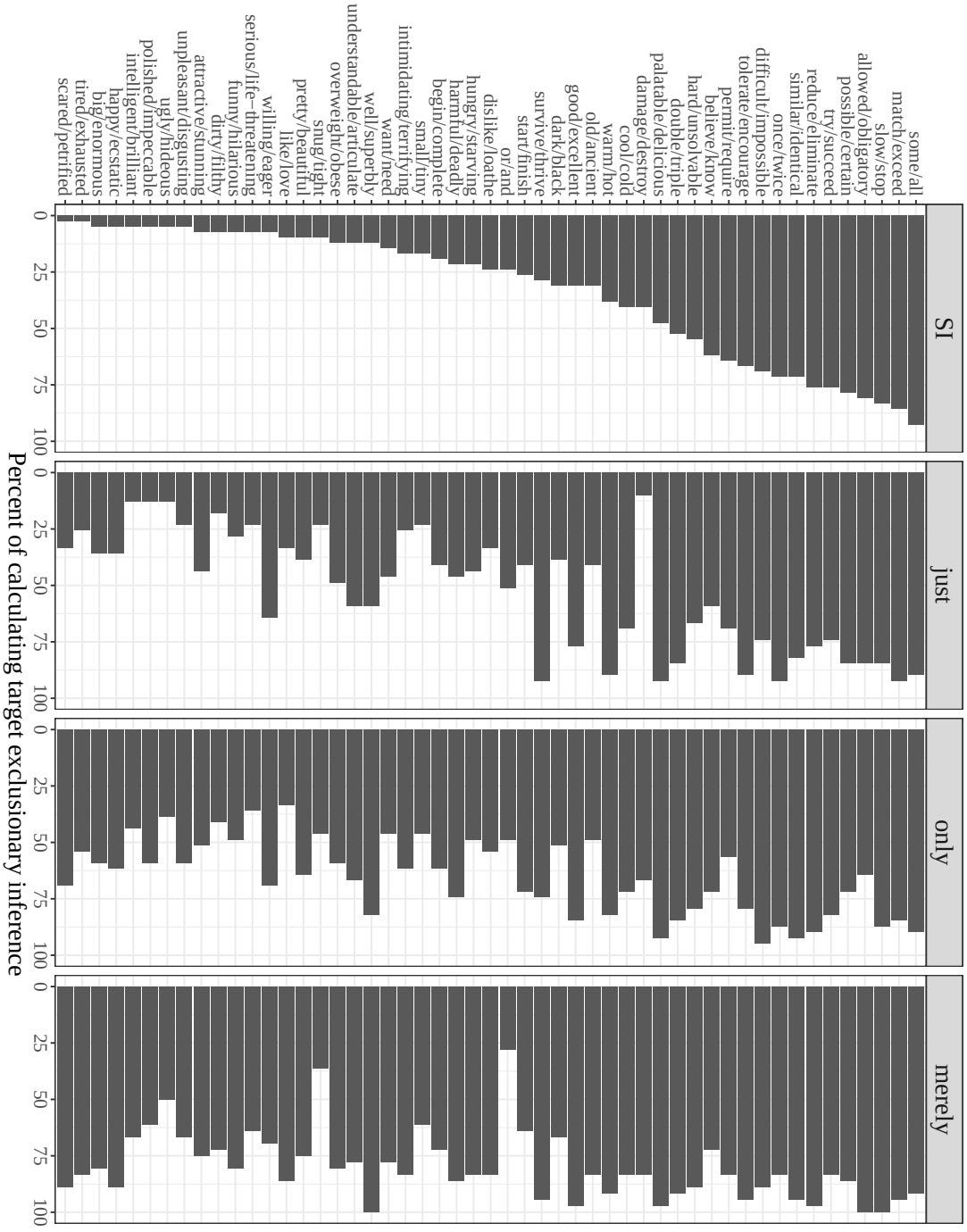


Figure 1: Results of Experiment 1 (*just*), Experiment 2 (*merely*), and Ronai and Xiang (2022)'s Experiment 1 (SI) and Experiment 3 (*only*).

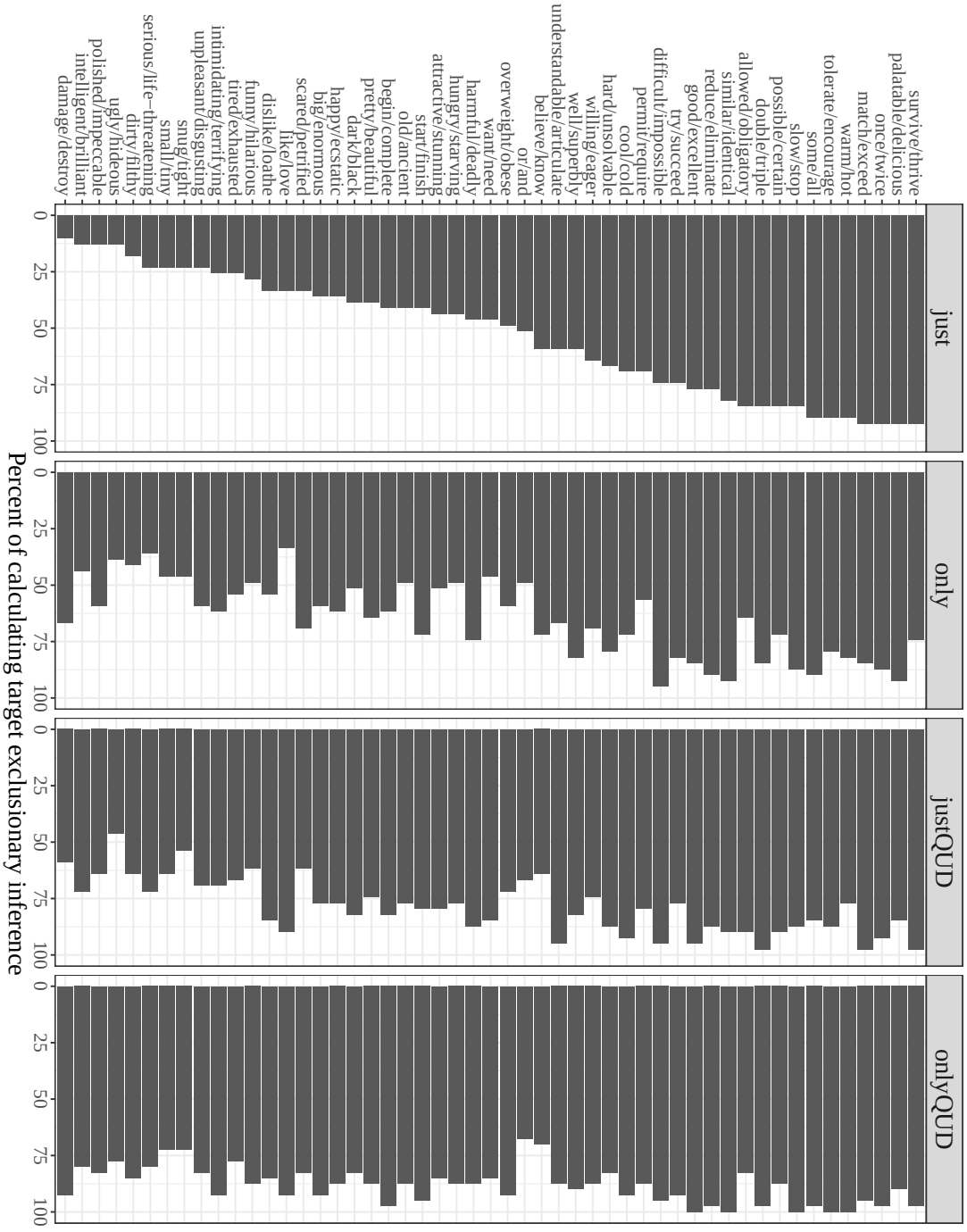


Figure 2: Results of Experiment 1 (*just*), Experiment 3 (*just* + QUD), and Ronai and Xiang (2022)'s Experiment 3 (*only*) and Experiment 4 (*only* + QUD).