
Ronai, Eszter & Fagen, Lucas. 2025. Experimental evidence for variation across exclusive modifiers. *Glossa: a journal of general linguistics* 10(1). pp. 1–22. DOI: https://doi.org/10.16995/glossa.16797

Open Library of Humanities

# Experimental evidence for variation across exclusive modifiers

**Eszter Ronai,** Northwestern University, US, ronai@northwestern.edu

**Lucas Fagen,** The University of Chicago, US, lfagen@uchicago.edu

This paper is an investigation of variation across the English exclusive modifiers *only*, *just*, and *merely*—a domain that has received ample attention in the theoretical literature but has thus far not been subjected to experimental testing. Using scalar diversity as a testing ground, we report on two experiments: Experiment 1 tests the robustness of exclusionary inference calculation (e.g., *merely intelligent* → *not brilliant*), and Experiment 2 directly tests whether a rank-order (e.g., *not brilliant*) or complement-exclusion (e.g., *not ambitious*) reading is preferred with different exclusives. Our findings reveal that 1) *just* excludes less robustly than the other two exclusives, and 2) while *only* allows both complement-exclusion and rank-order readings, *just* has a weak and *merely* a strong preference for rank-order ones. These results bear on previous theoretical observations about exclusives, and they are also informative about the robust by-scale variation in inference calculation, i.e., scalar diversity. Lastly, we argue that the methodological success of our experiments opens up avenues for further examination of more precise predictions of competing theoretical accounts.

# 1 Introduction

This squib experimentally investigates lexical semantic differences between the English exclusive modifiers *only*, *just*, and *merely*. Exclusives, as shown in (1), form a lexical class.

(1) a. Mary **only** ate [the cookies]$_F$.
b. Mary **just** ate [the cookies]$_F$.
c. Mary **merely** ate [the cookies]$_F$.

Sentences with exclusives typically convey that some proposition is true—the *prejacent*, usually treated as a presupposition since Horn (1969)—as in (2a), and that alternatives to the prejacent are false, as in (2b).

(2) a. Mary ate the cookies.
b. Mary did not eat alternatives to the cookies.

Beyond their common core meaning, exclusives vary considerably across different parameters (Coppock & Beaver 2014). For example, individual exclusives exhibit significant flexibility in the syntactic category and semantic type of what they modify. Exclusives also impose different restrictions on the alternatives to the prejacent and how they can be ordered along differently structured scales (Horn 2000; Klinedinst 2005; Beaver & Clark 2008; Fagen 2025). Additionally, some exclusives vary more widely than others. *Just* especially has a range of "noncanonical" (Warstadt 2020: p. 374) readings that *only* and *merely* lack, leading some researchers (Warstadt 2020; Beltrama 2021) to posit variation in the strength of the exclusive operation.

At the same time, variation between exclusives can be quite subtle and difficult to pin down. Such effects may emerge more starkly across items in an experimental setting than via intuition alone, and are therefore worth testing directly. Here, we present (to our knowledge) the first experimental assessment of this domain, using the scalar diversity phenomenon (van Tiel et al. 2016) as a testing ground. To preview our results, we find that 1) the likelihood of an exclusive inference is lower with *just* than with *only* and *merely*, and 2) the three exclusives vary in the extent to which they allow the exclusion of different kinds of alternatives: *merely* strongly prefers "rank-order" scales, *just* has a similar but weaker preference, while *only* freely allows both "rank-order" and "complement-exclusion" scales. We argue that the methodological success of our experiments also opens up avenues for further examination of the influence of semantic and pragmatic factors on the different parameters of variation.

The paper is structured as follows. We first review existing theoretical proposals about the semantics of exclusives as well as work on scalar diversity. We then present two experiments that compare how robustly different exclusives exclude alternatives, what type of alternatives they

tend to exclude, and how this interacts with scalar diversity. Finally, we discuss the theoretical and methodological implications of our results.

## 1.1 Scale structure

Exclusives can impose different orderings on the alternatives they exclude. (3) can be interpreted exhaustively, as conveying that the student possesses absolutely no relevant properties other than *intelligent* (3a). Following Coppock & Beaver (2014), we'll call this the *complement-exclusion* reading. (3) can also be interpreted as only excluding alternatives ordered higher than the prejacent along some pragmatically determined dimension (3b). We'll call this the *rank-order* reading.

(3) The student is **only** [intelligent]$_F$.

a. → The student has no other relevant properties (not curious, charming, etc).

b. → The student is not brilliant.

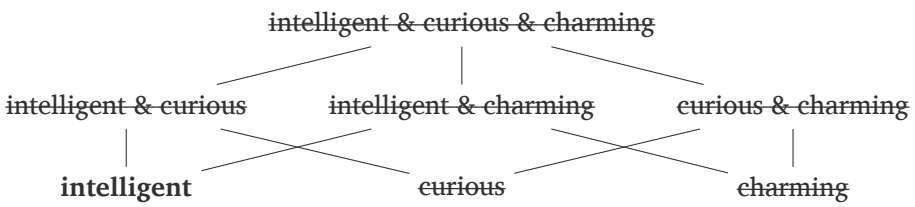
These readings can be given a uniform semantic analysis (Bonomi & Casalegno 1993; van Rooij 2002; Klinedinst 2005; Beaver & Clark 2008; Coppock & Beaver 2014). For example, Coppock & Beaver (2014) propose a unified typology of exclusives using two operators MIN and MAX. The presuppositional content of exclusives is specified in terms of MIN, and the assertive content is specified in terms of MAX. MIN and MAX apply to the prejacent and an ordered set of alternatives, which are modeled as answers to the current Question Under Discussion (CQ). Following Beaver & Clark's (2008) implementation of QUDs as ordered sets of propositions, or "scales", the ordering relation $\geq$ is retrievable from the CQ. MIN contributes existential quantification over the alternatives: some alternative ranked at least as high as the prejacent is true (4a). MAX contributes universal quantification: the prejacent is ranked at least as high as all true alternatives. The inference that the prejacent itself is true results from the combined effect of MIN and MAX.

(4) a. $\text{MIN}(p) = \lambda w.\exists q_{\in \text{CQ}}[q(w) \wedge q \geq p]$

b. $\text{MAX}(p) = \lambda w.\forall q_{\in \text{CQ}}[q(w) \rightarrow p \geq q]$

Different readings (e.g., (3a) vs. (3b)) result from variation in the ordering relation and the structure of the alternative set (Coppock & Beaver 2014). Complement-exclusion scales can be represented as a semilattice closed under conjunction (5), in which the higher-ranked alternatives entail the lower ones. In (5), the prejacent is *The student is intelligent* (abbreviated as *intelligent*). Application of MAX excludes all alternatives that are not ranked lower than the prejacent, resulting in the reading that the student is intelligent and has no other relevant properties. We indicate this in (5) by **bolding** the prejacent and ~~crossing out~~ the excluded alternatives.

(5)




Rank-order scales impose an ordering on atomic alternatives (6). We consider Horn scales like <*clever, intelligent, brilliant*> paradigmatic examples of rank-order scales, as they involve atomic alternatives that stand in an entailment relation. However, since rank-order scales are not closed under conjunction and lack the semilattice structure of (5), the alternatives can also be logically independent or mutually exclusive (e.g., <*assistant professor, associate professor, full professor*>), and can in principle be ordered by a wider range of relations. In this paper we will mostly focus on Horn scales.

(6)


Different exclusives are compatible with differently structured scales, but there is debate as to how freely exclusives actually vary in this regard. Horn (2000) argues from data involving projection (7) that *only* sentences presuppose the prejacent itself, while *just* sentences presuppose a lower bound on the prejacent. Since the *only* sentences presuppose the prejacent, it projects under negation, which leads to oddness when the alternatives are mutually exclusive. By contrast, no oddness is perceived with the *just* sentences.

(7) (Horn 2000: p. 150–151, ex. 7c, 9b)

a. They're not just/?only engaged, they're married.

b. I didn't just/?only get a B on the test, I got an A.

Translated into Coppock & Beaver's (2014) scalar framework, Horn's argument amounts to the claim that *only* selects for complement-exclusion scales and *just* selects for rank-order scales. Since stronger alternatives on a complement-exclusion scale entail the prejacent, negating the exclusive content preserves the inference to the prejacent. With rank-order scales, the alternatives do not need to stand in an entailment relation, so the inference to the prejacent can disappear under negation.

The generalization that *only* selects for complement-exclusion and *just* for rank-order scales might be too strong: some speakers can access the rank-order reading with *only* in examples like (7). Coppock & Beaver reinterpret the contrast between *just* and *only* in (7) as indicating

that exclusives have weaker pragmatic preferences for different scales: "*only* prefers entailment scales [i.e., complement-exclusion scales] and *just* has a slight preference for non-entailment scales [i.e., rank-order scales]" (Coppock & Beaver 2014: p. 425), without committing to a pragmatic analysis of what these preferences are or what factors influence them. The authors analyze *merely* as selecting an "evaluative" rank-order scale whose alternatives are ordered according to what the speaker considers good or bad. This is motivated by examples like (8), in which the higher-ranked alternatives are considered "better" (for further discussion of evaluativity effects with *merely*, see Wiegand, 2018; Windhearn, 2021; as well as Orenstein & Greenberg, 2010 on the Hebrew exclusive *stam*).

(8) How can people be happy or satisfied with merely the norm?
(Coppock & Beaver 2014: p. 424, ex. 180)

The goal of both experiments in this paper is to provide a quantitative experimental assessment of these claimed pragmatic preferences in scale structure.

## 1.2 Strength of exclusion

No theory of exclusives modeled after Horn's (1969) analysis of *only*, Coppock & Beaver's MIN/MAX entry schema included, can straightforwardly account for the *just* examples in (9). As the paraphrases show, these examples do not obviously communicate the exclusion of alternatives in the same way as the examples in (1).

(9) a. I was sitting there and the lamp **just** broke! (Wiegand 2018: ex. 1b)
*Paraphrase: Nothing caused the lamp to break.*
b. The lights in this place **just** turn off and on. (Warstadt 2020: ex. 1a)
*Paraphrase: The lights turn off and on for no reason.*
c. I'm not mad at you. I'm **just** mad. (Warstadt 2020: ex. 11a)
*Paraphrase: I'm not mad at anyone in particular.*
d. The essay is **just** perfect. (Beltrama 2021: ex. 1a)
*Paraphrase: The essay is perfect and nothing more needs to be said.*

One way that prior literature has accounted for such uses of *just* is by positing that it differs from other exclusives in (what we refer to as) its strength of exclusion. For example, Warstadt (2020) proposes an entry for *just* that differs from *only* and other exclusive modifiers in two places. First, *just* excludes answers to "potential" questions, or "intuitively possible future QUDs" (p. 373, see also Onea 2016), rather than the current QUD. Second and crucially for our purposes, Warstadt argues that a unified account of *just*'s various readings requires relaxing the truth-conditional status of the exclusive operation. He proposes a distinction between "strong"

exclusives which declare alternative propositions false, and "weak" exclusives which declare them unassertable.[1]

On Warstadt's analysis, the semantic contribution of *just* in, e.g., (9b) is as follows. A speaker who asserts that the lights turn off and on might anticipate the addressee asking why this is. *Just* marks the answers to this potential question as unassertable, before the addressee can ask why, "thus preventing the addressee from asking a useless question" (Warstadt 2020: p. 373). However, *just* can also target the current QUD. In this case, the claim that other alternative answers are unassertable, together with the presupposed truth of the prejacent, leads to a meaning very similar to what a canonical exclusive entry would have delivered, the primary difference being one of strength: a speaker who uses *just* is refusing to commit to either the truth or falsity of alternatives, unlike a speaker who uses *only* to mark alternatives as false. Warstadt analyzes the full range of *just*'s readings as involving this weaker exclusive operation—in contrast with *only*, which Warstadt analyzes as a strong exclusive.

Other accounts of non-canonical readings with *just* have posited other sources of variation in the set of alternatives *just* can operate on, suggesting that it can target: covert modifiers with trivial semantic content (Wiegand 2018; Windhearn 2021), alternative scale granularities (Thomas & Deo 2020), or metalinguistic alternatives at the speech act level (Beltrama 2021).

Altogether, whether an exclusive is weak or strong arises as a subtle pragmatic effect. Experiment 1 therefore tests *just*'s claimed "weakness", or more generally, the possibility that it allows for non-canonically exclusive readings.

## 1.3 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; Gotzner et al. 2018; Sun et al. 2018). A classic example of SI is (10): upon encountering an utterance containing *some*, hearers compute the negation of its stronger scalar alternative *all* (Grice 1967; Horn 1972).

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

Similarly to <*some, all*>, e.g. <*intelligent, brilliant*> also forms a scale: an utterance containing *intelligent* can lead to the SI *not brilliant* (11). But variation exists across different scales: the SI in (10) arises much more robustly than the one in (11).

[1] See also Beltrama (2021), who proposes that emphatic *just* as in (9d) declares alternatives not "worthy of assertion" (p. 347).

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

Scalar diversity persists even in the presence of exclusives. Ronai & Xiang (2024) found that even when sentences such as (10)–(11) 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. Ronai & Xiang 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). By experimentally comparing *only* to other exclusives, potentially less flexible in scale structure preference, we will also be able to test Ronai & Xiang's hypothesis.

As mentioned, although Horn scales like <*some, all*> and <*intelligent, brilliant*> stand in an entailment relation, we class these as rank-order scales because the structure of the alternative set does not bottom out in logically independent "atomic" propositions; that is, the scale structure is as in (6), not (5). Throughout this paper we use the terms "complement-exclusion" and "rank-order" to characterize scale structure rather than the precise relations used to order alternatives.

## 2 Experiment 1: Inference task

To investigate the lexical semantics of exclusive modifiers, we compare how they affect scalar diversity. We replicate Ronai & Xiang's (2024) experiment testing *only*, and add as comparison manipulations with *just* and *merely*. This will allow us to test two predictions made by the theoretical literature on exclusives—reviewed above in Sections 1.1–1.2. First, since our experiment will specifically test rank-order alternatives, Coppock & Beaver's (2014) account predicts that alternative exclusion will be more robust with *merely* than with *only*, since these authors analyze the former as preferring rank-order readings. Second, the claims by Warstadt (2020) about strength of exclusion predict that participants will be less likely to exclude alternatives with *just* than with *only*. If *just* marks alternatives as unassertable rather than false, participants should be more reluctant to infer that alternatives are false. It is possible that Warstadt's unified analysis is on the wrong track and *just* is instead lexically ambiguous between exclusive and nonexclusive readings. In that case, we still predict lower rates of exclusion, on the assumption that participants who interpret *just* non-exclusively on some trials will be less likely to commit to excluding the stronger alternative.

## 2.1 Participants

120 monolingual native speakers of American English participated in the experiment (40 in each of the between-participants conditions), which was administered on the web-based PCIbex platform (Zehr & Schwarz 2018). Participants were recruited on Prolific and compensated $2. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Participants were excluded for making more than 3 mistakes on the 7 attention check items. After exclusions, data from 111 participants was analyzed (37 for *only*, 39 for *just*, and 35 for *merely*.)

## 2.2 Materials and procedure

The experiment used a two-alternative forced choice inference task, identical to Ronai & Xiang's experiments (see also i.a. van Tiel et al. 2016). On each trial, participants saw a target sentence uttered by Mary, which contained an exclusive (e.g., *The student is {just/only/merely} intelligent*), and they were asked whether Mary thinks that the rank-order alternative is not true (e.g., the student is not brilliant). **Figure 1** shows an example with *just*. Here, a "Yes" response indexes that the participant has calculated the exclusionary (*not brilliant*) inference, while selecting "No" suggests that the inference was not calculated. The experiment tested three different exclusives—*only*, *just*, and *merely*—as a between-participants manipulation.

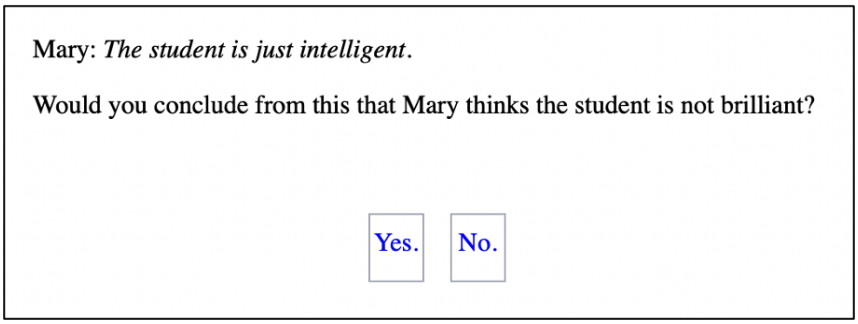



**Figure 1:** Example experimental trial from Experiment 1 (*just* condition).

We tested 51 lexical scales as critical items, a subset of those used by Ronai & Xiang (2024). Of their 60 items, we removed those incompatible with *just* or *merely*. This amounted to removing a subset of the adverbial scales, e.g., <*mostly, entirely*> (*#Peter's answers were merely [mostly]$_F$ wrong.*); <*primarily, exclusively*> (*#The residents are merely [primarily]$_F$ Greek.*); or <*probably, necessarily*> (*#A delay will merely [probably]$_F$ occur.*). An additional modification was made to the verbal scales in the *just* condition: wherever a temporal interpretation of *just* would have been possible, we modified the tense of the carrier sentence to rule this out. For example, for <*survive, thrive*>, *The plant only/merely survived* was changed to *The plant is just surviving*. Similarly, for

*<reduce, eliminate>*, *The city only/merely reduced waste* was modified to *The city will just reduce waste*.[2]

In addition to the critical items, the experiment tested 7 fillers that served as attention checks. The fillers, adapted from Ronai & Xiang (2024) and van Tiel et al. (2016), contained two antonyms (*wide → not narrow*) and therefore had an unambiguous "Yes" answer. The experiment began with 2 practice items.

## 2.3 Results

**Figure 2** shows the results of Experiment 1. For the statistical analysis, 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 predictions were made for *just* vs. *only* (strength of exclusion) and *merely* vs. *only* (scale structure bias), 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 $= -0.63$, SE $= 0.27$, $z = -2.33$, $p < 0.05$), as well as significantly higher rates with *merely* than with *only* (Estimate $= 1$, SE $= 0.3$, $z = 3.4$, $p < 0.001$). Averaged over the 51 different scales, the target exclusionary inference was calculated at the rate of 52.9% with *just*, 63.2% with *only*, and 80.2% with *merely*.

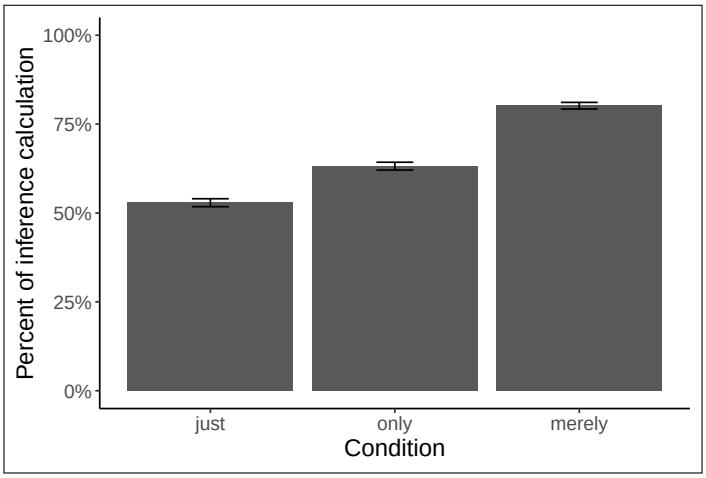



**Figure 2:** Results of Experiment 1: *just*, *only*, *merely*. The y axis shows the percent of calculating the target exclusionary inference, i.e., the mean percent of "Yes" responses in the inference task. Error bars represent standard error.

[2] An additional concern might be whether *just* can be interpreted emphatically. We think this is unlikely, since in our experiment *just* always modified a weaker scale-mate (*just intelligent*), and for an emphatic use it would typically need to modify an extreme adjective (Morzycki 2012; Beltrama 2021). To the extent that extreme adjectives occurred in our experiment, they would have been in the task question ("Would you conclude from… not brilliant?"), not modified by *just*.

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 also conducted a replication of Ronai & Xiang's (2024) SI experiment.[3] We found an average rate of 34.3% SI calculation; **Figure 3** shows the SI results together with the three exclusives, also visualizing the by-item variation. An additional statistical model compared the findings of this experiment to *just*. The fixed effects predictor was sum-coded (SI: −0.5 and *just*: 0.5). Inference rates were found to be significantly higher with *just* than in the case of SI (Estimate = 1.29, SE = 0.27, $z = 4.8$, $p < 0.001$). This confirms that alternative exclusion with all three exclusive modifiers is stronger than alternative exclusion via SI.

One may wonder whether different lexical scales interact differently with the two tested parameters of variation, scale structure bias and strength of exclusion. In order to check whether the relative order of the 51 items remained consistent across manipulations, we calculated rank-order correlations using Kendall's $\tau_B$, and found significant by-item correlations between conditions. As SI rates increase, so do inference rates with *just* ($\tau_B = 0.67$, $p < 0.001$); as rates with *just* increase, so do rates with *only* ($\tau_B = 0.64$, $p < 0.001$); and rates with *merely* are also correlated with *only* ($\tau_B = 0.39$, $p < 0.001$).[4] In other words, we found that items that lead to low SI rates also lead to relatively low rates of inference calculation with exclusives, even as the overall rate of inference calculation goes up with an exclusive as compared to SI. To give an example, <*small*, *tiny*> led to SIs at a rate of 7.9%, and with exclusives, it also continued to trigger inference calculation at a rate lower than the average for that exclusive: 23.1% with *just*, 43.2% with *only* and 61.1% with *merely*.

## 2.4 Discussion

Both predictions made by the relevant theoretical accounts regarding strength of exclusion and scale structure bias were borne out by the results. First, Experiment 1 found lower rates of exclusionary inference calculation—that is, lower rates of "Yes" responses in the inference task—with *just* than with *only*. This is consistent with Warstadt's (2020) proposal that *just* is a weak exclusive, while *only* is a strong exclusive. Or alternatively and more generally, this finding confirms that *just* is not always canonically exclusive; if (some) participants accessed a non-exclusive reading, that could have lowered the overall observed rate of exclusion. Second, Experiment 1 found higher inference rates with *merely* than with *only*. Since all our items tested rank-order alternatives, this strongly supports Coppock & Beaver's (2014) claim that while *only*

[3] The SI replication was identical to the main Experiment 1, with the exception that target sentences (i.e., Mary's utterances) did not contain an exclusive. 40 participants were recruited; 1 was excluded for failing attention checks and 1 for being bilingual.

[4] Though still highly significant, the *only-merely* correlation is less strong than the other two pairs, i.e., the coefficient is lower. This is expected, since with *merely*, inference rates are more uniform, closer to ceiling, and therefore the relative difference between any two lexical scales is much smaller.

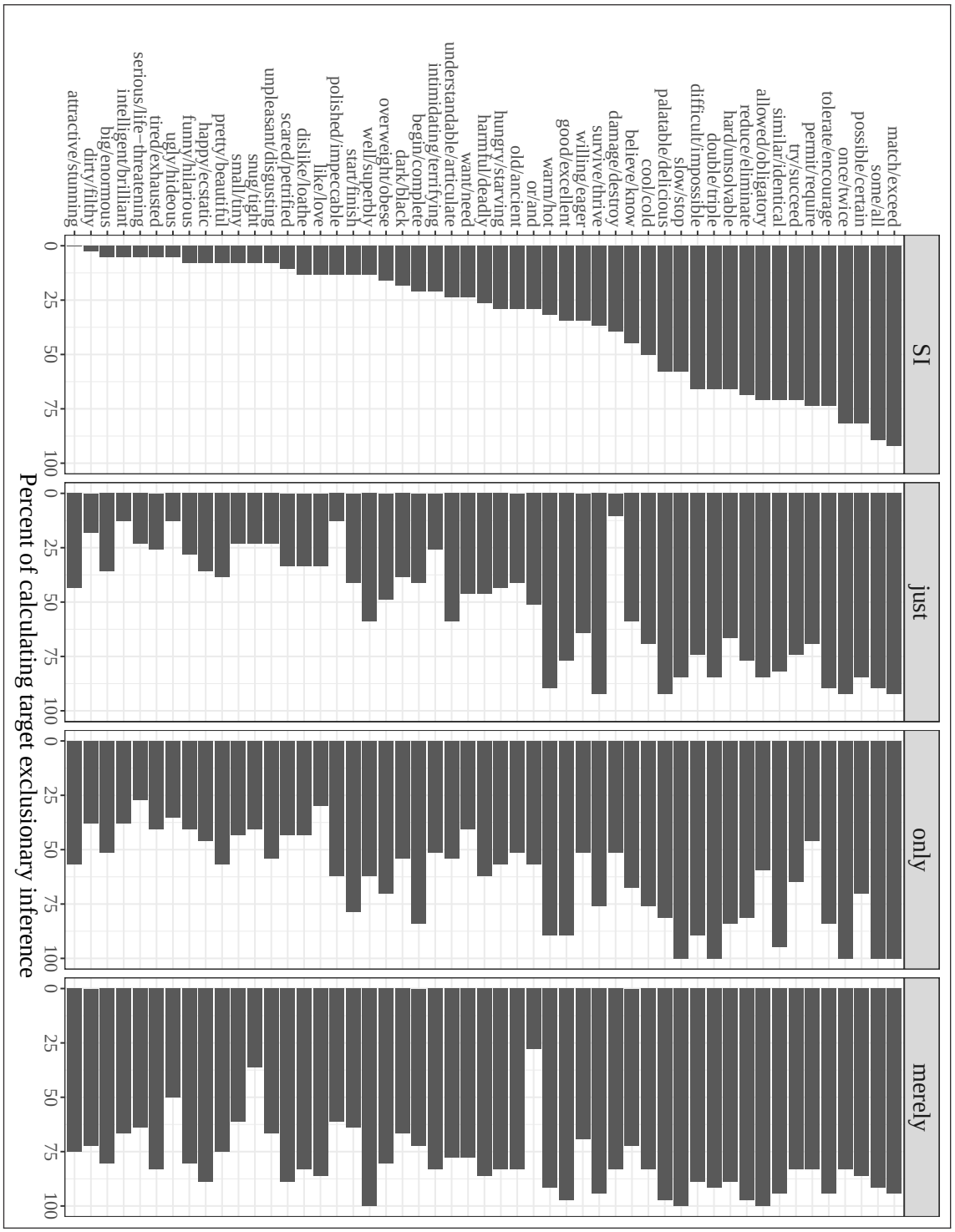



**Figure 3:** Results of Experiment 1: SI, *just*, *only*, *merely*. The percent of exclusionary inference calculation corresponds to the percent of "Yes" responses in the inference task.

allows both complement-exclusion and rank-order readings, *merely* prefers rank-order ones. These results also support Ronai & Xiang's (2024) hypothesis that when interpreting target sentences with *only*, participants are split between rank-order and complement-exclusion readings. When a rank-order bias is introduced by *merely*, the stronger scalar term in the "Would you conclude from this…?" task question is more unambiguously understood as one of the salient alternatives, which leads to an increase in calculating the target inference.

While we have found evidence that *only* and *merely* differ in scale structure, with the latter more strongly requiring rank-order alternatives, Experiment 1 does not allow us to straightforwardly infer anything about the (potential) scale structure bias of *just*. Coppock & Beaver (2014) argue that, like *merely*, *just* also has a(n albeit weaker) preference for rank-order readings (p. 425, see also Horn 2000). This would predict a higher rate of inference calculation with *just*, as compared to *only*. However, *just*'s comparatively weak strength of exclusion pushes in the other direction, which as we have seen has resulted in overall lower inference rates. Therefore, the current results leave open two possibilities: either *just* does not favor rank-order scales, contra the theoretical literature, or it does, but that preference in the Experiment 1 design was counteracted by its weak exclusive status. Experiment 2 was designed to adjudicate between these two possibilities.

## 3 Experiment 2: Scale choice

In Experiment 1, we were not able to draw conclusions about *just*'s scale structure bias, since, as our findings confirmed, it excludes relatively less robustly than the other two modifiers. To address this, in Experiment 2, we employ a novel implementation of the two-alternative forced choice task, which allows us to probe what kind of alternative hearers prefer to exclude in a setting where one must be excluded. We test this preference for all three exclusives. As before, Coppock & Beaver's account predicts that the choice of rank-order alternatives will be more robust with *merely* than with *only*. As for *just*, both Horn (2000) and Coppock & Beaver (2014) predict that *just* will prefer rank-order scales, albeit to different degrees: for Horn, *just* is completely restricted to rank-order scales, whereas for Coppock & Beaver this preference is a weaker pragmatic effect.

### 3.1 Participants

120 monolingual native speakers of American English participated in the experiment (40 in each of the between-participants conditions), administered on PCIbex. Participant recruitment and screening was identical to Experiment 1; compensation was $2 or $2.40 depending on time. 5 participants were removed from the *merely* condition and 1 participant from the *just* condition for making 3 or more mistakes on the 5 attention check items. Data from the remaining 114 participants is reported below.

## 3.2 Materials and procedure

Experiment 2 ruled out a non-exclusive interpretation by employing a two-alternative forced choice task where participants had to choose between two alternatives as the target for exclusion. Specifically, they were presented with a target utterance by Mary, e.g., *That student was just intelligent*, which was followed by "Mary meant that the student was not BLANK". Participants had to choose between a rank-order and a complement-exclusion alternative to fill in the blank.[5] **Figure 4** shows an example trial with *just*.

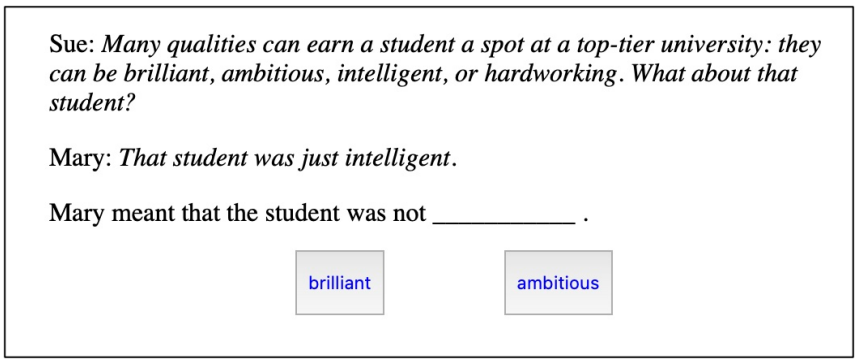



**Figure 4:** Example experimental trial from Experiment 2 (*just* condition).

Target inference-triggering sentences occurred in the context of Sue's preceding utterance. The context provided by Sue always explicitly introduced alternatives, including the focus associate (here, *intelligent*), the rank-order alternative (*brilliant)*, the complement-exclusion alternative (*ambitious*), and a fourth alternative (*hardworking*) that was not probed in the forced choice task. Our task was designed to make both the rank-order and complement-exclusion alternatives salient and relevant, with the expectation that with salience and relevance controlled for, participants' responses would directly reflect the scale structure bias contributed by different exclusives.

This design choice was inspired by previous experimental studies on the processing of focus with *only*, which have successfully used discourse contexts to introduce relevant alternatives. (12) illustrates this with an example from Hoeks (2023)—see also Washburn et al. (2011); Kim et al. (2015); Fraundorf et al. (2010) for similar designs.

[5] Given the assumptions about scale structure we make in (5), one might object that alternatives based on single properties like *ambitious* are not really complement-exclusion alternatives, and that the actual complement-exclusion alternative in a trial like that shown in **Figure 4** would be the conjunction of an independent property with the focus associate, in this case *intelligent* & *ambitious*. Since the exclusion of the stronger conjunction entails the negation of the atomic proposition in context, we take it that participants who select *ambitious* in this task have still accessed the complement-exclusion reading.

(12) The tourist asked for a variety of items, like some cheese and milk. There was already an ashtray on the table. When the waiter returned, he remembered to bring only milk but no cheese to the table where the tourist was seated.

Inherent to the goals of our experiment is that the context sentence uttered by Sue would have to introduce two alternatives that are ranked relative to one another (*intelligent* and *brilliant* in **Figure 4**), in addition to an alternative that is not inherently related to either (i.e., the complement-exclusion one, *ambitious* in **Figure 4**). In order to mitigate against potential unnaturalness arising from this design, we took the following steps. First, as mentioned, a fourth alternative which can be classified as complement-exclusion (*hardworking* in **Figure 4**) was also included, in order to make sure that the target complement-exclusion alternative (*ambitious*) was not overly salient as being the exception. Second, to ensure a natural reading experience throughout the experiment and to prevent participants from noticing ordering patterns, the order of the four alternatives was randomized on every trial. Analysis of the Experiment 2 data suggests that the design did not introduce a bias towards one kind of alternative as compared to Experiment 1 (see footnote 6).

The basic frame of the inference-triggering sentences was kept similar to Experiment 1 (and Ronai & Xiang 2024). The focus associates (i.e., weaker scalar terms embedded under an exclusive) and rank-order alternatives were identical to Experiment 1. Complement-exclusion alternatives were generated for each item, using corpus searching (for strings such as *only X but not Y*, with *X* as the focus associate) and the intuitions of the authors. The fourth alternative was similarly generated by author intuition.

Experiment 2 tested 48 items. 3 items were removed from the item set of Experiment 1 because we were unable to generate a suitable complement-exclusion alternative (<*some, all*>, <*or, and*>, <*once, twice*>). We also included 3 practice items and 5 fillers to serve as catch trials. In filler items, one of the two forced-choice alternatives was the focus associate itself (e.g., *The street is just wide*—Mary meant that the street is not: wide vs. tree-lined) and hence unambiguously the incorrect choice.

## 3.3 Results

**Figure 5** shows the results of Experiment 2. Averaged over the 48 different scales, the rank-order alternative (as opposed to the complement-exclusion alternative) was chosen 56.8% of the time with *only*,[6] 65.6% with *just*, and 80.8% with *merely*. For the statistical analysis, a logistic mixed

[6] A reviewer notes that the likelihood of rank-order choices with *only* was lower in Experiment 2 (56.8%) than the corresponding likelihood of "Yes" responses in Experiment 1 (63.2%) and wonders whether the contexts provided in Experiment 2 made participants less able to access rank-order interpretations. To explore this possibility, we fit a logistic mixed effects regression model to the combined *only* data from Experiments 1 and 2 (restricted to the 48 scales tested in both). The model included random intercepts for participants and random slopes and intercepts for items and

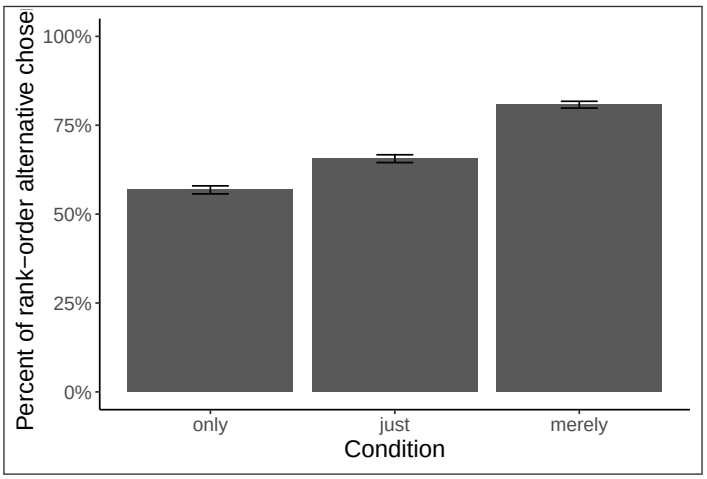



**Figure 5:** Results of Experiment 2: Mean percent of choosing rank-order (as opposed to complement-exclusion) alternative with *only* vs. *just* vs. *merely*. Error bars represent standard error.

effects regression model was fit (lme4). We compared Response (rank-order vs. complement-exclusion choice) for each exclusive to chance, i.e., to 50% rank-order/complement-exclusion choice. Chance level reflects the hypothetical whereby an exclusive has no constraints on its scale structure, and freely allows either type of alternative to be excluded; in this case, we assume participants would choose one over the other with 50% likelihood. The model included random intercepts for participants and random slopes and intercepts for items. We found that *just* (Estimate $= 0.98$, SE $= 0.26$, $z = 3.78$, $p < 0.001$) and *merely* (Estimate $= 2.01$, SE $= 0.26$, $z = 7.88$, $p < 0.001$) differed significantly from chance level by producing a higher rate of rank-order choice. On the other hand, results with *only* did not significantly differ from chance (Estimate $= 0.44$, SE $= 0.27$, $z = 1.65$, $p = 0.1$).

In order to directly compare the behavior of the three different exclusives to one another, we fit an additional model that predicted Response (rank-order vs. complement-exclusion choice) as a function of Exclusive (*just* vs. *only* vs. *merely*). Here, the Exclusive predictor was treatment-coded, with *just* serving as the reference level. We found that *merely* produced significantly higher rates of rank-order choice than *just* (Estimate $= 1.03$, SE $= 0.32$, $z = 3.21$, $p < 0.01$). The *only* vs. *just* difference (i.e., *only* producing fewer rank-order choices) was smaller and failed to reach significance (Estimate $= -0.54$, SE $= 0.3$, $z = -1.81$, $p = 0.07$). As visual inspection of **Figure 6** shows, this was likely due to by-item variation.

predicted Response by Experiment (sum-coded, Experiment 1: −0.5 and Experiment 2: 0.5). The model revealed that difference between the likelihood of "Yes" responses in Experiment 1 and of rank-order choices in Experiment 2 was not significant (Estimate $= -0.33$, SE $= 0.31$, $z = -1.08$, $p = 0.28$). This suggests that the two different experimental paradigms did not affect participants' ability to assign rank-order interpretations.

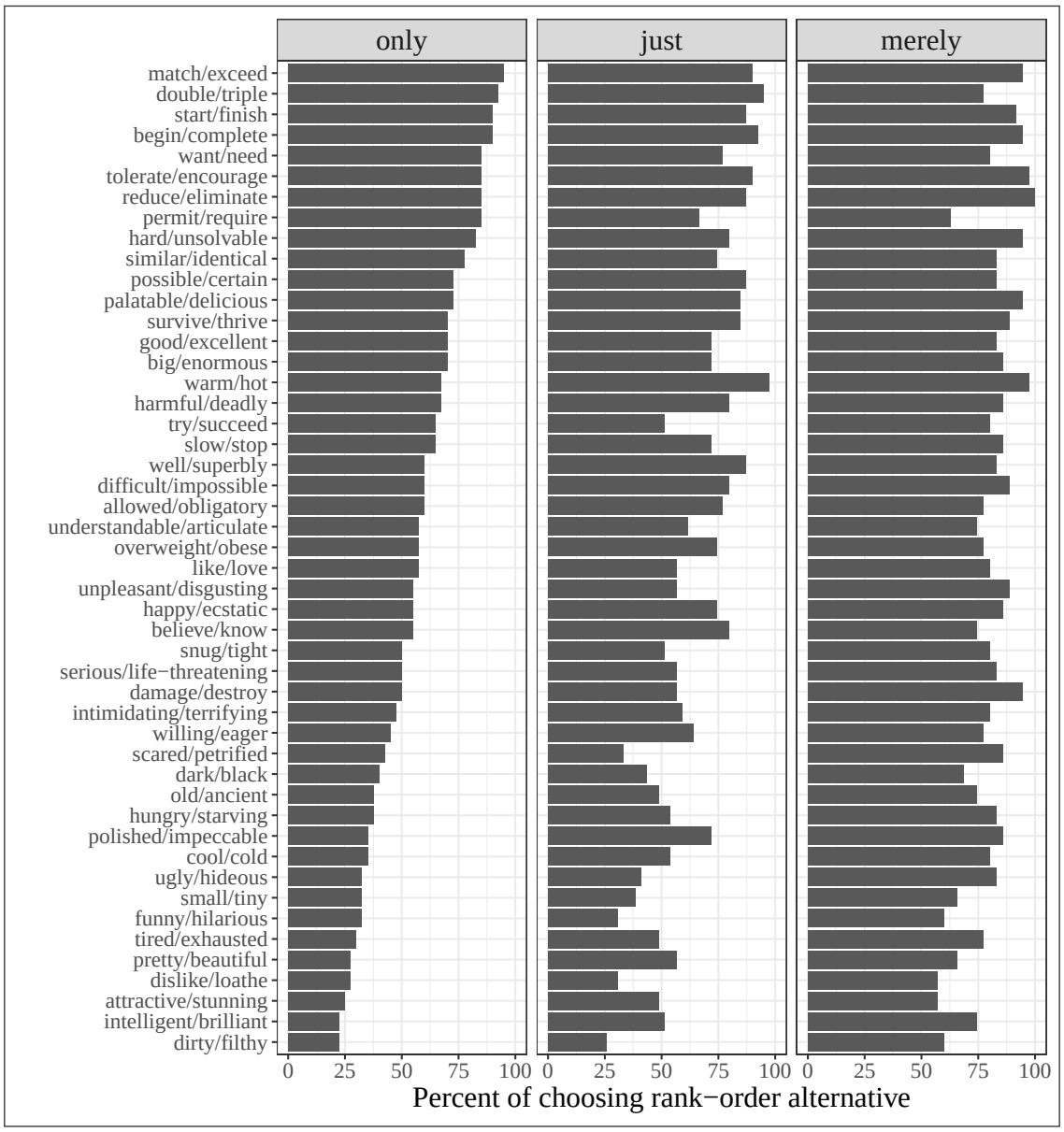



**Figure 6:** Results of Experiment 2: By-scale variation in choosing rank-order alternative.

## 3.4 Discussion

By probing whether participants took the target sentence to exclude a rank-order or a complement-exclusion alternative, instead of probing whether they calculated the target exclusionary inference, we were able to more clearly isolate scale structure bias in Experiment 2. We found that *merely* and *just* are clearly different from chance level in showing a preference for the exclusion of rank-order alternatives. *Merely* strongly preferring rank-order exclusion successfully replicates

Experiment 1, while the *just* finding goes beyond the previous experiment with the confound of a non-exclusive reading now ruled out. Unlike the other two exclusives, *only* was found not to be different from chance level, suggesting that it allows both types of alternatives (more) freely.

Experiment 2's findings are in line with Coppock & Beaver's (2014) observations about the scale structure bias of *merely* and *just*. These authors argue that *merely* requires rank-order alternatives, while *just*'s preference for them is only slight. Our finding that *merely* and *just* patterned significantly differently from each other constitutes strong support for this claim. The results of Experiment 2, on the other hand, are less compatible with Horn (2000), whose proposal leads to the expectation that *just* selects for rank-order scales—our data suggests that *just*'s preference is more subtle than that.

## 4 General discussion

This squib presented novel experimental evidence testing variation across exclusive modifiers in English. In this domain, previous claims about available readings and speaker preferences between readings were based on particular, often idiosyncratic examples (e.g., (9)). We sought evidence for graded pragmatic preferences between readings in a controlled experimental setting. Our results provided support for two claims from the theoretical literature: that exclusives vary in scale structure bias (with *merely* and *just* both preferring rank-order readings, the former more so than the latter) and in strength of exclusion (with *just* excluding less robustly).

Our *merely* results straightforwardly support Coppock & Beaver's (2014) claim that *merely* is restricted to ("evaluative") rank-order scales. Since all our Experiment 1 items tested rank-order alternatives, the stronger scalar alternative was more frequently included in the alternative set given an exclusive restricted to rank-order readings, leading to higher rates of target inference calculation compared to *only*. Experiment 1's results also support Ronai & Xiang's (2024) hypothesis that participants in their *only* experiment—which we successfully replicated—were split between rank-order and complement-exclusion readings. This additionally confirms Coppock & Beaver's claim that both readings are indeed available for *only*, a claim at odds with Horn (2000), who instead analyzes *only* as selecting complement-exclusion scales. The *merely* findings were replicated in Experiment 2. The design of Experiment 2, which ruled out the possibility of non-exclusion, additionally allowed us to draw conclusions about the scale structure bias of *just*. We found that it indeed favors rank-order scales, although not to the extent that *merely* does—a pattern more in line with Coppock & Beaver (2014) than Horn (2000). Our overall results support Coppock & Beaver's view of scale structure bias as a matter of pragmatic preference, as well as their specific claims about which exclusives favor which scales.

Our *just* results in Experiment 1 verify the recurrent intuition in the literature that *just* excludes less robustly than *only*. Different analyses of the apparent strength difference have been proposed, several of which are compatible with our results. If *just* excludes alternatives on the

basis of pragmatic assertability rather than truth (i.e., "weak" exclusion), as argued by Warstadt (2020) and Beltrama (2021), participants could have been more reluctant to conclusively reject alternatives with *just* compared to *only*. An equally plausible, if less interesting interpretation would attribute the lower inference rates to lexical ambiguity: participants excluded the stronger scalar alternative when *just* was interpreted exclusively and not otherwise.

To our knowledge, this squib reports on the first experimental investigation of variation across exclusive modifiers. We hope that the success of the experimental methods employed here, including the novel design of Experiment 2, opens up possibilities for future research which can address this variation. We found strong evidence that different exclusives favor different sorts of scales, and we succeeded in precisely quantifying the strength of these preferences by exclusive. Prior theoretical literature (e.g., Horn 2000; Coppock & Beaver 2014) has observed these preferences on a descriptive level, but to our knowledge there has been little work aimed at uncovering their source. If scale structure bias is a pragmatic phenomenon that is not fully determined by lexical semantics, it is possible that exclusives compete with each other: the use of one exclusive over another could trigger defeasible pragmatic inferences about what sorts of alternatives the speaker intended to exclude, and how these alternatives were ordered. For example, the use of *only*, which is lexically unrestricted with respect to scale structure, over an exclusive like *just* or *merely*, which prefer rank-order scales, could lead to a preference for the complement-exclusion reading with *only* in some contexts. Since the speaker did not use an exclusive that would have explicitly enforced the rank-order reading, a pragmatic listener might reason that a complement-exclusion reading must have been intended. Independent lexical semantic differences between exclusives could also make one or another exclusive a better "fit" for certain scales over others; this is arguably the case with *merely*, whose evaluative nature—which is in principle independent of scale structure—could underlie its preference for rank-order scales. These questions are difficult to pursue on an intuitionistic level. We are optimistic that the experimental methods employed in this paper—which make it possible to systematically manipulate relevance, salience, and other pragmatic factors—can therefore provide a basis for such investigations.

Lastly, though in our experiments scalar diversity was a testing ground, not the main object of study, the results are nonetheless informative with respect to this phenomenon. In particular, we found significant by-item rank-order correlations across all conditions in Experiment 1. That is, if we order different lexical scales (e.g., <*intelligent, brilliant*> vs. <*allowed, obligatory*>) based on their likelihood of leading to an exclusionary inference, this order remains consistent across experiments: from SI to different exclusive modifiers. This is despite the fact that as inference rates increase across the board (reaching on average 80.2% with *merely*), there is necessarily less variation across lexical scales, and their relative ranking becomes less meaningful. This finding points to the importance of lexico-semantic factors in the scalar diversity phenomenon.

Specifically, prior work has shown that properties of lexical scales such as the distinctness of the weaker and stronger scale-mates (van Tiel et al. 2016), their semantic relatedness (Westera & Boleda 2020), the expectedness of the stronger alternative (Hu et al. 2023), or adjectival polarity and extremeness (Gotzner et al. 2018; Beltrama & Xiang 2013) can predict a scale's likelihood of leading to SIs. In this paper, we did not directly test whether these properties are also significant predictors of the variation found in our *just*, *only*, and *merely* results. Nonetheless, it seems likely that since such properties reliably predict variation in SI rates, and—as evidenced by the significant by-item correlations—overall the same variation arises with exclusives, factors like distinctness, etc. could underlie the variation in the exclusives data as well. Additionally, exclusives can be thought of as a probe for the alternative expectedness predictor in particular. As Ronai & Xiang (2024) have argued based on their *only* findings, scales that do not robustly trigger the calculation of the target exclusionary inference with an exclusive are likely those where the stronger alternative is not expected (Hu et al. 2023) or available (van Tiel et al. 2016). This is because exclusives like *only* semantically encode the exclusion of an alternative; therefore, if a participant nonetheless responds "No" in the inference task, that can be attributed to their inability to take the provided stronger alternative (e.g., *brilliant* given *only intelligent*) to be a relevant alternative to the weaker term. The argument also applies to our Experiment 1, and our rank-order correlation findings suggest that it can be extended to the exclusives *just* and *merely* as well.

As such, the findings reported in this paper shed light not only on parameters of variation across exclusives, but also on the variation across lexical scales in their propensity to lead to exclusionary inferences.

## Data availability

Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/ktrzm/?view_only=36c09d6b218d4f73b00cc01c860184e4.

## Ethics and consent

The study reported as Experiment 1 in this paper was approved by the University of Chicago Institutional Review Board (#IRB20-2038-CR003). The study reported as Experiment 2 was granted exemption by the Northwestern University Institutional Review Board (#STU00220908) due to being low risk and involving only tests, surveys, interviews, observation, or benign behavioral interventions.

## Funding information

This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312.

## Acknowledgements

We thank editor Lyn Tieu, three anonymous reviewers, Itamar Francez, Chris Kennedy, and Michael Tabatowski as well as audiences at the 23rd Amsterdam Colloquium, the 97th Annual Meeting of the LSA, the UChicago LEAP workshop and the Northwestern Experimental Meaning Group for helpful feedback and discussion. All remaining errors are our own.

## Competing interests

The authors have no competing interests to declare.

## Authors' contributions

The authors contributed equally to this work and are listed in reverse alphabetical order.

## References

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Beaver, David & Clark, Brady. 2008. *Sense and sensitivity: how focus determines meaning.* Chichester: Wiley-Blackwell. DOI: https://doi.org/10.1002/9781444304176

Beltrama, Andrea. 2021. Just perfect, simply the best: an analysis of emphatic exclusion. *Linguistics and Philosophy* 45. 321–364. DOI: https://doi.org/10.1007/s10988-021-09326-x

Beltrama, Andrea & Xiang, Ming. 2013. Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In Chemla, Emmanuel & Homer, Vincent & Winterstein, Grégoire (eds.), *Proceedings of Sinn und Bedeutung 17*, 81–98.

Bonomi, Andrea & Casalegno, Paolo. 1993. *Only*: association with focus in event semantics. *Natural Language Semantics* 2. 1–45. DOI: https://doi.org/10.1007/BF01255430

Coppock, Elizabeth & Beaver, David. 2014. Principles of the exclusive muddle. *Journal of Semantics* 31(3). 371–432. DOI: https://doi.org/10.1093/jos/fft007

Fagen, Lucas. 2025. Exclusives and scale structure. Ms., The University of Chicago.

Fraundorf, Scott H. & Watson, Duane G. & Benjamin, Aaron S. 2010. Recognition memory reveals just how contrastive contrastive accenting really is. *Journal of Memory and Language* 63(3). 367–386. DOI: https://doi.org/10.1016/j.jml.2010.06.004

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

Grice, Herbert Paul. 1967. Logic and Conversation. In Grice, Paul (ed.), *Studies in the Way of Words*, 41–58. Harvard University Press.

Hoeks, Morwenna. 2023. *Comprehending focus/representing contrast*: University of California, Santa Cruz dissertation.

Horn, Laurence. 1969. A presuppositional analysis of only and even. In *Proceedings of CLS 5*.

Horn, Laurence. 1972. *On the semantic properties of logical operators in English*: UCLA dissertation.

Horn, Laurence. 2000. Pick a theory (not just any theory). In Horn, Laurence & Kato, Yasuhiko (eds.), *Negation and polarity: syntactic and semantic perspectives*, Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198238744.001.0001

Hu, Jennifer & Levy, Roger & Degen, Judith & Schuster, Sebastian. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*. To appear. DOI: https://doi.org/10.1162/tacl_a_00579

Kim, Christina S. & Gunlogson, Christine & Tanenhaus, Michael K. & Runner, Jeffrey T. 2015. Context-driven expectations about focus alternatives. *Cognition* 139. 28–49. DOI: https://doi.org/10.1016/j.cognition.2015.02.009

Klinedinst, Nathan. 2005. *Scales and* only: UCLA MA thesis.

Morzycki, Marcin. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language & Linguistic Theory* 30(2). 567–609. DOI: https://doi.org/10.1007/s11049-011-9162-0

Onea, Edgar. 2016. *Potential questions at the semantics-pragmatics interface*. Brill. DOI: https://doi.org/10.1163/9789004217935

Orenstein, Dina & Greenberg, Yael. 2010. The semantics and focus sensitivity of the Hebrew (unstressed) *stam*. In *Proceedings of IATL 26*.

Ronai, Eszter & Xiang, Ming. 2024. What could have been said? Alternatives and variability in pragmatic inferences. *Journal of Memory and Language* 136. 104507. DOI: https://doi.org/10.1016/j.jml.2024.104507

Sun, Chao & Tian, Ye & Breheny, Richard. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9. DOI: https://doi.org/10.3389/fpsyg.2018.02092

Thomas, William & Deo, Ashwini. 2020. The interaction of *just* with modified scalar predicates. In *Proceedings of Sinn und Bedeutung 24*.

van Rooij, Robert. 2002. Relevance only. In *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialogue*.

van Tiel, Bob & Miltenburg, Emiel Van & Zevakhina, Natalia & Geurts, Bart. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175.

Warstadt, Alex. 2020. "Just" don't ask: exclusives and potential questions. In Franke, Michael & Kompa, Nikola & Liu, Mingya & Mueller, Jutta L. & Schwab, Juliane (eds.), *Proceedings of Sinn und Bedeutung 24*, vol. 2. 373–390.

Washburn, Mary Byram & Kaiser, Elsi & Zubizarreta, Maria Luisa. 2011. Focus facilitation and non-associative sets. *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue* 94–102.

Westera, Matthijs & Boleda, Gemma. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung* 24(2). 439–454. DOI: https://doi.org/10.18148/sub/2020.v24i2.908

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

Windhearn, Mia. 2021. *Alternatives, exclusivity, and underspecification*. Cornell University dissertation.

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