2021. *Proc Ling Soc Amer* 6(1). 649–662. https://doi.org/10.3765/plsa.v6i1.5001.

Exploring the connection between Question Under Discussion and scalar diversity

Eszter Ronai & Ming Xiang*

Abstract. Previous research has revealed that different scalar expressions give rise to scalar inferences (SIs) at different rates. This variation has been termed *scalar diversity*. In this study, we investigate the role of Questions Under Discussion (QUDs) in explaining this variation in SI rates. Investigating 43 different scalar expressions, we first show that explicit QUDs robustly affect calculation rates: questions based on the stronger of two scalar terms lead to higher SI rates than questions promoting the weaker one. Second, we explore whether the likelihood of asking the stronger question in general (Question Availability) can explain the scalar diversity effect. Our results suggest that Question Availability is indeed a predictor of scalar diversity, but only for scales where both terms denote intervals (unbounded scales), and not for scales where the stronger member has a fixed meaning (bounded scales).

Keywords. experimental pragmatics; Question Under Discussion; scalar inference; scalar diversity

1. Background. In successful communication, comprehenders regularly infer meanings that go beyond what is literally, explicitly said by the speaker. A well-known example of this phenomenon is scalar inference (SI), illustrated in (1).

(1) Mary ate some of the cookies.
Literal meaning: Mary ate some, and possibly all, of the cookies.
Inference-enriched meaning: Mary ate some, but not all, of the cookies.

Upon encountering the sentence in (1), comprehenders reason about informationally stronger alternatives that could have been said in place of what was actually said. In particular, they are taken to reason that the lexical alternative *all*, or the alternative utterance *Mary ate all of the cookies* was also available to the speaker, and that because they chose not to utter that alternative, its negation can be inferred. Comprehenders thus arrive at the inference-enriched meaning: *some but not all* (Grice 1967).

1.1. SCALAR DIVERSITY. While much research has investigated the SI in (1), which is based on the <*some, all*> scale, there exist many other scales where a set of lexical items are are ordered with respect to each other in terms of their logical strength (Horn 1972). The example in (2), for instance, is based on the <*intelligent, brilliant*> scale.

(2) The student is intelligent.
Literal meaning: The student is intelligent, and possibly brilliant.
Inference-enriched meaning: The student is intelligent, but not brilliant.

Such examples, in principle, give rise to SI the same way as <*some, all*>, e.g. comprehenders, upon encountering *The student is intelligent*, will reason about *The student is brilliant* as a potential alternative and infer its negation. However, recent work has found that there is in fact

* For helpful discussion and feedback, we thank Andrea Beltrama, Chris Kennedy, Michael Tabatowski, and the audience at LSA 95. All mistakes and shortcomings are our own. Authors: Eszter Ronai, The University of Chicago (ronai@uchicago.edu) & Ming Xiang, The University of Chicago (mxiang@uchicago.edu).

considerable variation across different scales in the rates of SI calculation; for instance, the SI in (1) arises much more robustly than the one in (2) (van Tiel et al. 2016; see also Doran et al. 2009, 2012; Beltrama & Xiang 2013).

The question arises, then, how to capture this observed variation: can we identify some properties of different scales that influence how robustly they lead to SI calculation? Van Tiel et al. (2016) has found distinctness of the stronger and weaker scalar terms to be such a property; the more distinct the two terms are, the more reasonable it is to assume that the speaker should have used the stronger term if possible. In particular, the authors operationalized distinctness as semantic distance and boundedness. Semantic distance (measured via a rating task) revealed that the more distant a weak and a strong scalar terms are, the stronger the SI from the weak term is. This can be intuitively seen in the below example:

(3) a. Many of the senators voted against the bill.
    b. Most of the senators voted against the bill.
    c. All of the senators voted against the bill.

In this example, the sentence in (3-a) is more more likely to lead to the the negation of (3-c) as an SI than to the negation of (3-b), because on the <*many, most, all*> scale, *all* is more distant from *many* than *most* is (Horn 1972). Distinctness had a second component, boundedness, with the findings that bounded scales, i.e. scales where the stronger scalar term refers to an end point, lead to higher SI rates. We return to boundedness in the discussion of our own experimental findings.

Subsequent work has identified further predictors of SI rates across different scales. Sun et al. (2018) has found that local enrichability also contributes to scalar diversity: e.g., the higher the naturalness of a sentences such as *Mary ate all, so not some, of the cookies*, the higher the SI corresponding rate. This is because, as the authors argue, to be able to interpret such a sentence as natural, rather than contradictory, the scalar term *some* must locally be interpreted on its inference-enriched meaning (*some but not all*). Focusing exclusively on adjectival scales, Gotzner et al. (2018) have found that polarity, as well as extremeness, can predict the variation in SI rates: negative scales (<*bad, awful*>) were found to yield higher SI rates than positive ones (<*good, great*>), while scales with extreme adjectives (*excellent, huge*) as their stronger members were found to have lower SI rates —for findings regarding extremeness, see also Beltrama & Xiang (2013). As with boundedness, we will return to adjectival extremeness in the interpretation of our own findings in Section 5.

Importantly for the present study, despite existing work identifying some properties of scales that significantly predict the relevant SI rates, there is still a great deal of variance unaccounted for in the empirical results. Specifically, van Tiel et al. found that in their statistical analysis, semantic distance explained 10% of the observed variance, while boundedness explained only 3%. In Sun et al. (2018)'s study, 15% of the variance was explained by propensity for local enrichment, while Gotzner et al. (2018) found that extremeness explained 17% and polarity 5% of the variance in their data. Models that include multiple known predictors from different studies still fall short of explaining all of the variance in SI rates, with Sun et al. (2018) reporting that their best fitted model explained 63% of the variance, and Gotzner et al. (2018) reporting 62%. In other words, a lot of scalar diversity is still unexplained.

1.2. QUESTION UNDER DISCUSSION AND THE PRESENT STUDY. It has long been noted that discourse context, formalized e.g. as Questions Under Discussion (QUDs), can modulate the likelihood of SI calculation (see i.a. Kuppevelt, 1996). Indeed, existing experimental work has found that manipulating the QUD, via explicit questions or a background story, has a significant effect on SI rates. Consider, for instance, the examples in (4)-(5).

(4) A: Did Mary eat all of the cookies?
B: Mary ate some of the cookies.

(5) A: Did Mary eat any/some of the cookies?
B: Mary ate some of the cookies.

It is a robust empirical finding that B's utterance, *Mary ate some of the cookies*, gives rise to the familiar *some but not all* SI at a higher rate in (4) than in (5) —see i.a. Degen & Tanenhaus (2015); Ronai & Xiang (2020); Yang et al. (2018); Zondervan et al. (2008) for converging results. Such findings can be given an explanation along the following lines. In (5), unlike in (4), A's question suggests that they are only interested in whether Mary has eaten at least some of the cookies. There is therefore no particular reason to consider *Mary ate all of the cookies* as an alternative that B could have said. This intuition can be formalized, for instance, by reference to the Question-Answer Requirement (Hulsey et al. 2004). Existing work on the role of discourse context in modulating SI rates has largely concentrated on the <*some, all*> scale; though see Cummins & Rohde (2015), who tested a larger number of different scales, albeit manipulating QUD via focus intonation.

In previous work on (factors explaining) scalar diversity, stimulus sentences were presented in the absence of any context. This raises the possibility that there could be variation across scalar terms in what kind of QUD they most naturally bring to mind. More specifically, in this study we explore the hypothesis that scalar diversity, in the absence of an explicit QUD, arises (in part) due to the differential availability of a polar question containing the stronger scalar term from the scale. To understand the intuition behind this hypothesis, let us first consider the two dialogues in (6)-(7).

(6) A: Is the student brilliant?
B: She is intelligent.

(7) A: Is the student intelligent?
B: She is intelligent.

In what follows, we refer to questions such as the one in (6) as *strong-scalar questions*, and questions such as the one in (7) as *weak-scalar questions*. Assuming that the <*intelligent, brilliant*> scale patterns similarly to the <*some, all*> scale, we predict higher rates of SI calculation in (6) than in (7). Building on this prediction, our main hypothesis is that when the sentence *She is intelligent* occurs without explicit discourse context, the likelihood of calculating the *intelligent but not brilliant* SI is a function of how likely comprehenders are to consider QUDs such as the one in (6), as opposed to the one in (7). In other words, the more likely a question such as *Is the student brilliant?* (the strong-scalar questions) is, the higher the rate of SI calculation from the corresponding statement will be. In what follows, we refer to this likelihood of asking the strong-scalar question as *Question Availability*, and explore whether variation in Question Availability across different scales can explain variation in SI rates across

those scales, that is, scalar diversity.

We conducted three experiments to explore our hypothesis. We first replicate van Tiel et al. (2016)'s experiment and main finding of scalar diversity (Experiment 1). In Experiment 2, we test whether an explicit QUD, i.e. an overt question in a dialogue context, has an effect on SI rates across a large number of different scales. In Experiment 3, we operationalize Question Availability and explore whether it predicts the rates of SI calculation from Experiment 1.

2. Experiment 1. Experiment 1 was a replication of van Tiel et al. (2016). We used an inference task to measure the rate of SI calculation for 43 different lexical scales.

2.1. PARTICIPANTS. 40 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Amazon Mechanical Turk and compensated $2.50. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. 3 participants were removed from analysis because the background questionnaire revealed that they were bilinguals; data from 37 participants is reported below.

2.2. TASK, MATERIALS AND PROCEDURE. We used an inference task to investigate the likelihood of deriving an SI from 43 different scales. Participants were presented with a sentence such as "Mary: *The student is intelligent.*" and were asked the question "Would you conclude from this that, according to Mary, the student is not brilliant?". They responded by clicking "Yes" or "No". Figure 1 shows an example trial item.

A "Yes" answer indicates that the participant has calculated the relevant SI (*intelligent* → *not brilliant*), while a "No" answer indicates that the participant has not calculated the SI, i.e. they are interpreting *intelligent* as meaning *intelligent and possibly brilliant*.

Our experimental materials used the 43 lexical scales from van Tiel et al. (2016)'s Experiment 2. Every participant saw each scale only once. Scales occurred in one of three carrier statements, which were varied between-participants; (8) shows the three carrier statements for the <*intelligent, brilliant*> scale.

(8) a. The assistant is intelligent.
b. The professor is intelligent.
c. The student is intelligent.

The materials were slightly modified from van Tiel et al. (2016) in that the distal demonstrative was always changed to the proximal demonstrative to increase naturalness, e.g. *That plant*

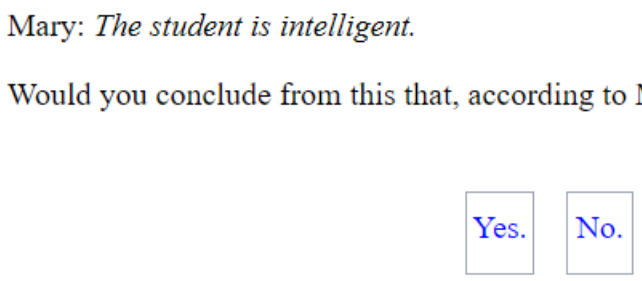

Figure 1. Example experimental trial from Experiment 1

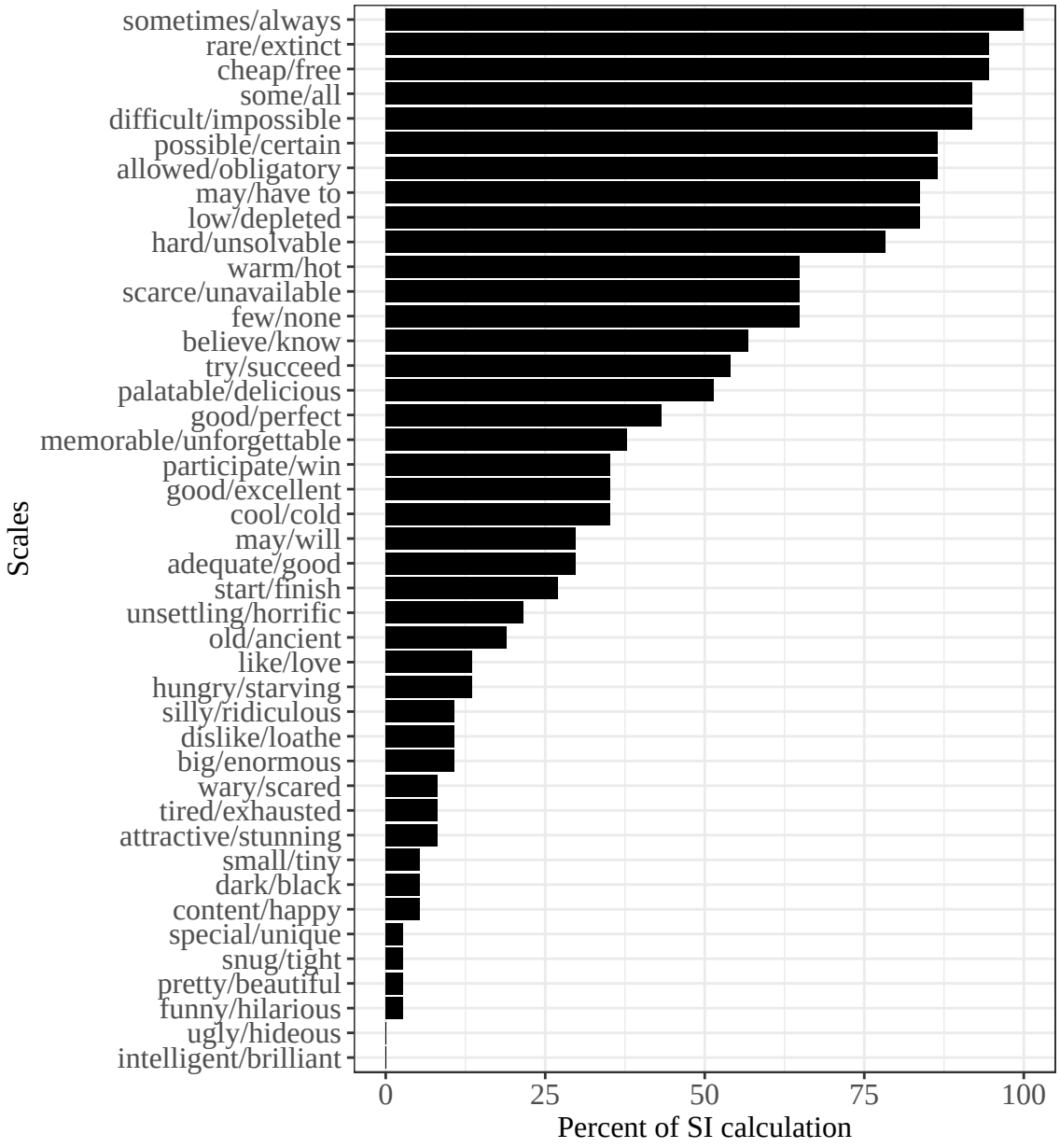

Figure 2. Results of Experiment 1: SI Rate for 43 different scales

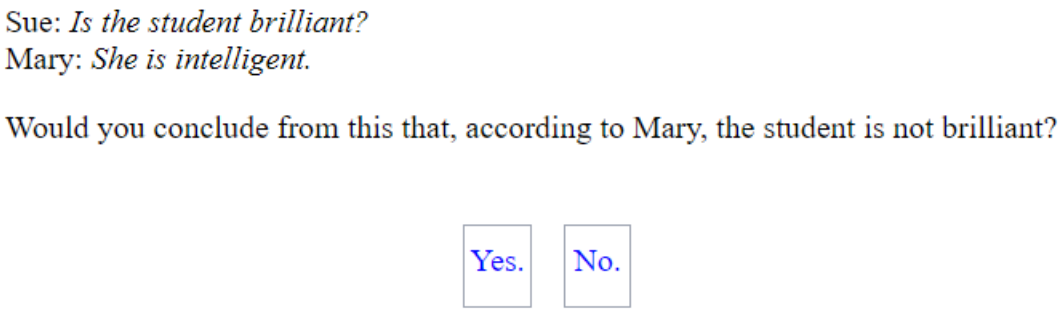

Figure 3. Example experimental trial from Experiment 2

is rare was changed to *This plant is rare*. 7 filler items were also included, which contained two terms that are either in an entailment relation (*dead* → *not alive*), or unrelated (*wide* → *not colorful*). Given that they had a clear, correct "Yes" or "No" answer, filler items were included to serve as catch trials. The experiment began with 2 practice trials to familiarize participants with the task; following that, each participant saw 50 trials. The experiment was administered in a Latin Square design.

2.3. PREDICTION. Given consistent findings of scalar diversity in existing literature, we predict robust variation across the 43 different scales in how likely they are to lead to SI calculation. That is, we predict that the percentage of "Yes" vs. "No" responses in the inference task of Experiment 1 will vary substantially from scale to scale.

2.4. RESULTS AND DISCUSSION. Figure 2 shows the results of Experiment 1. Percent of SI calculation corresponds to the the proportion of "Yes" responses. As is evident from this figure, we found considerable variation among critical items, with positive responses, i.e. rate of SI calculation, ranging along a continuum from 0% (for <*intelligent, brilliant*> and <*ugly, hideous*>) to 100% (for <*sometimes, always*>). This result thus successfully replicates the scalar diversity effect: different scalar expressions yielded widely different rates of SI.

3. Experiment 2. Experiment 2 tested the effect of QUDs on the rate of SI calculation for a large number of different scales. We used the same task as Experiment 1, but the potentially SI-triggering sentences now served as answers in a dialogue context, and the preceding question was manipulated.

3.1. PARTICIPANTS. 40 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Amazon Mechanical Turk and compensated $2.50. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Data from all 40 participants is reported below.

3.2. TASK, MATERIALS AND PROCEDURE. Experiment 2 employed the same task as Experiment 1, but with the important addition of a two-condition Question manipulation: each statement from Experiment 1 was embedded in a dialogue context. Specifically, Mary's statement was preceded either by a question containing the stronger scalar term, or by a question containing the weaker scalar term. That is, for the <*intelligent, brilliant*> scale, the weak-scalar question was *Is the student intelligent?*, while the strong-scalar question was *Is the student brilliant?*. Figure 3 shows an example trial item, from the strong-scalar Question condition.

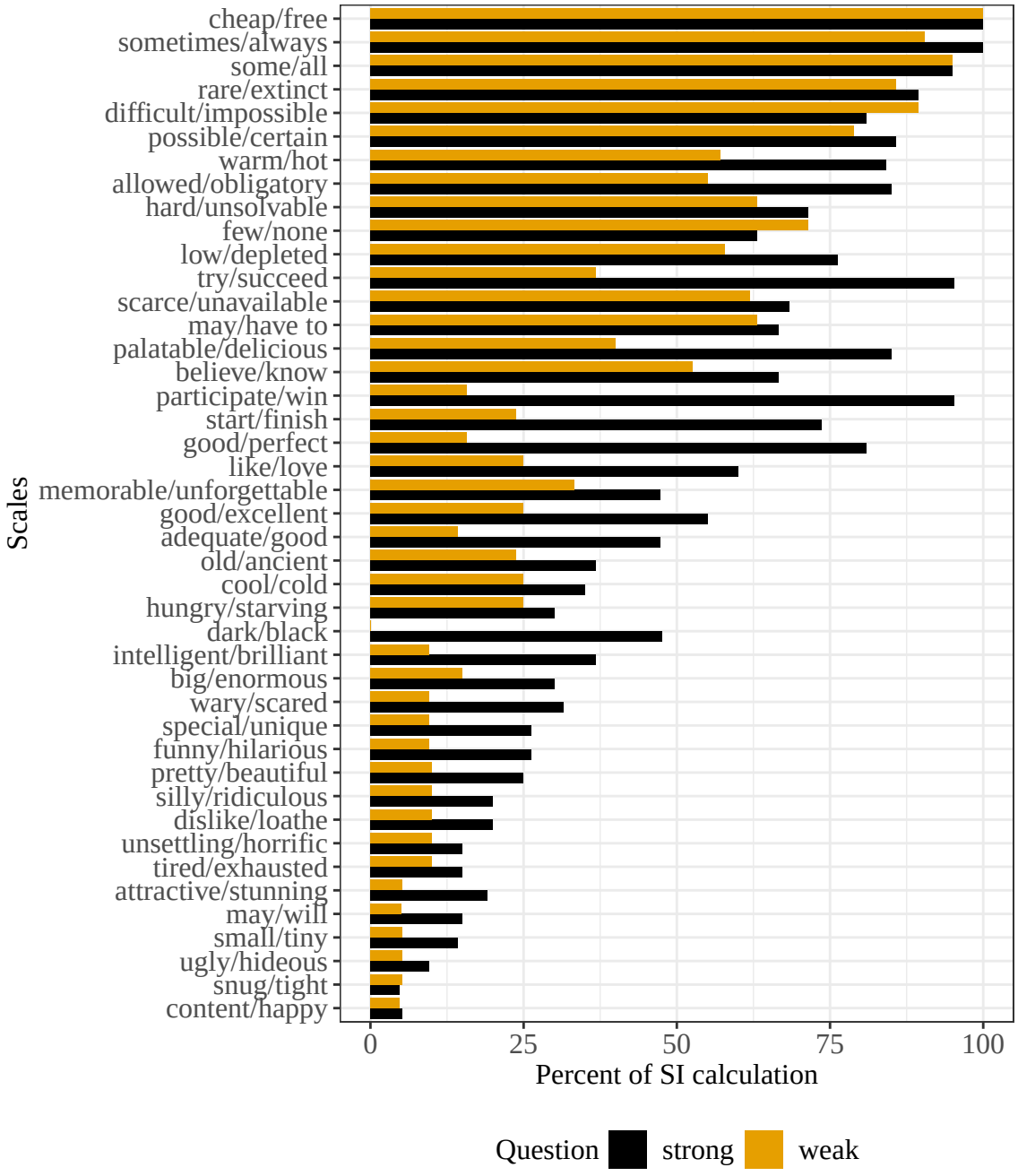

Figure 4. Results of Experiment 2: SI Rate for 43 different scales, by Question condition.

The potentially SI-triggering statements, i.e. Mary's answers, were slightly modified to ensure dialogue coherence, e.g. *The student is intelligent* was changed to *She is intelligent*. Apart from this modification, and the addition of the Question manipulation, the materials and procedure of Experiment 2 were identical to Experiment 1.

3.3. PREDICTION. Existing work has found an effect of QUDs such that, for instance, a question containing *all* leads to a higher level of SI calculation on the <*some, all*> scale than a question containing *some* —see Section 1.2. We predict this modulating effect of QUDs, operationalized in Experiment 2 as overt questions, to extend to other scales as well. Specifically, we predict an overall effect of the Question manipulation, such that strong-scalar questions lead to higher SI rates than weak-scalar questions for the 43 scales tested.

Additionally, we predict that, if the scalar diversity effect is, at least in part, due to the differential availability of the strong-scalar question across scales, then explicitly providing that strong-scalar question in a dialogue context may reduce (or eliminate) scalar diversity. In other words, even though sentences containing e.g. *intelligent* and *some* lead to different SI calculation rates in the absence of any context, they might lead to similar rates of SI calculation when they follow a question containing *brilliant* and *all*, respectively.

3.4. RESULTS AND DISCUSSION. Figure 4 shows the results of Experiment 2. For each scale, we calculated the percentage of SI calculation (=SI Rate). For the statistical analysis, a linear regression model was fitted (lm in R), predicting SI Rate by Question. The Question variable was sum-coded before analysis. We found a significant effect of the Question manipulation ($\beta = 9.03$, $t = 2.76$, $p < 0.01$). This effect is driven by strong-scalar questions leading to higher SI rates than weak-scalar questions; across the board, more SIs were derived when the preceding question contained the stronger scalar term than when it contained the weaker one. However, the scalar diversity effect is still present: we did not find that embedding SI-triggering sentences under the strong-scalar questions reduces the variation in SI rates.

4. Experiment 3. Experiment 3 operationalized Question Availability and tested our main hypothesis: that the likelihood of SI calculation from a given scale (e.g. <*intelligent, brilliant*>) can, in part, be explained by the likelihood of a QUD that contains the stronger term from that scale (*brilliant*).

4.1. PARTICIPANTS. 40 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Amazon Mechanical Turk and compensated $2.50. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. 4 participants were removed from analysis because the background questionnaire revealed that they were bilinguals, and 1 additional participant was removed based on having a reaction time shorter than 300ms on the majority of the trials, suggesting that they were not paying attention to the task. Data from 35 participants is reported below.

4.2. TASK, MATERIALS AND PROCEDURE. Experiment 3 employed a forced choice task. Participants had to choose which of two polar questions they would be more likely to ask: the one containing the stronger or the weaker scalar term from the given scale. Participants responded by clicking on one of the two questions. Figure 5 shows an example trial item.

This task is motivated by the assumption that how likely participants are to choose a question in a given context (e.g. when talking about a student) is an index of how available that

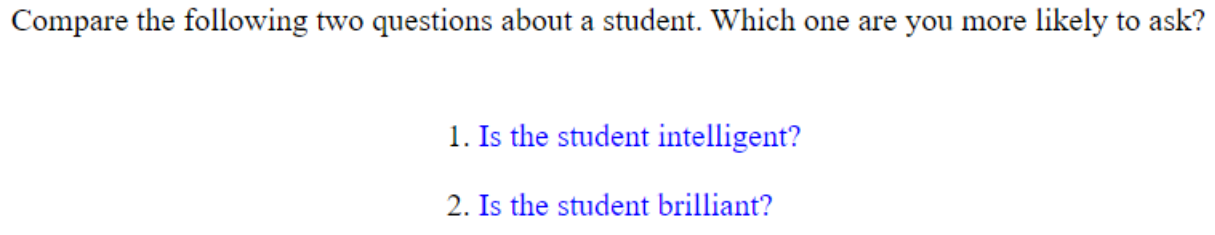

Figure 5. Example experimental trial from Experiment 3

question is to them in the absence of any specific discourse context. In other words, the more participants in Experiment 3 prefer the question *Is the student brilliant?* over *Is the student intelligent?*, the more likely they are to reason about that strong-scalar question when no discourse is provided.

Experiment 3 used the same 43 scales (in three different carrier statements), 7 filler items, and 2 practice items as the previous two experiments.

4.3. PREDICTION. Under our hypothesis, the results from Experiment 3 (henceforth Question Availability) should predict scalar diversity, i.e. the variation in SI calculation rates from Experiment 1 —the more preferred the strong-scalar question is in Experiment 3, the higher the SI rate should be for that scale in Experiment 1.

4.4. RESULTS AND DISCUSSION. For each scale, we took the percentage of SI calculation from Experiment 1 (=SI Rate). From Experiment 3, we calculated the percentage of choosing the strong-scalar question for each scale (=Question Availability). For the statistical analysis, a linear regression model was fitted (lm in R), predicting SI Rate by Question Availability. We found that Question Availability was not an overall predictor of SI Rate ($\beta = 0.03$, $t = 0.17$, $p = 0.87$).

Next we conducted a post-hoc analysis that takes into account the boundedness of the scales—see Figure 6. Van Tiel et al. (2016) define a scale as bounded if the stronger scalar denotes an endpoint: <*some, all*> is therefore an example of a bounded scale, while <*intelligent, brilliant*> is unbounded. We took van Tiel et al. (2016)'s categorizations of scales as either bounded or unbounded and included it in our analysis as a predictor. We fit a linear regression model predicting SI Rate by Question Availability, Boundedness, and their interaction. The variable of Boundedness was sum-coded before analysis. We found a significant effect of Boundedness ($\beta = 33$, $t = 5.9$, $p < 0.001$). This effect is driven by bounded scales producing significantly higher SI rates than unbounded ones—a replication of van Tiel et al. (2016). Crucially, we also found a significant interaction of Question Availability with Boundedness ($\beta = -0.25$, $t = -2.3$, $p < 0.05$).

Following this, we analyzed bounded and unbounded scales in separate regression models, predicting SI Rate by Question Availability. For unbounded scales, Question Availability showed a strong numerical trend ($\beta = 0.17$, $t = 1.86$, $p < 0.08$) in predicting SI Rate. As can be seen in Figure 6, for unbounded scales, the more likely participants were to choose the strong-scalar question (*Is the student brilliant?*), the higher the rate of calculating the relevant SI (*intelligent→not brilliant*) was. For bounded scales, however, Question Availability had no effect ($\beta = -0.33$, $t = -1.54$, $p = 0.14$).

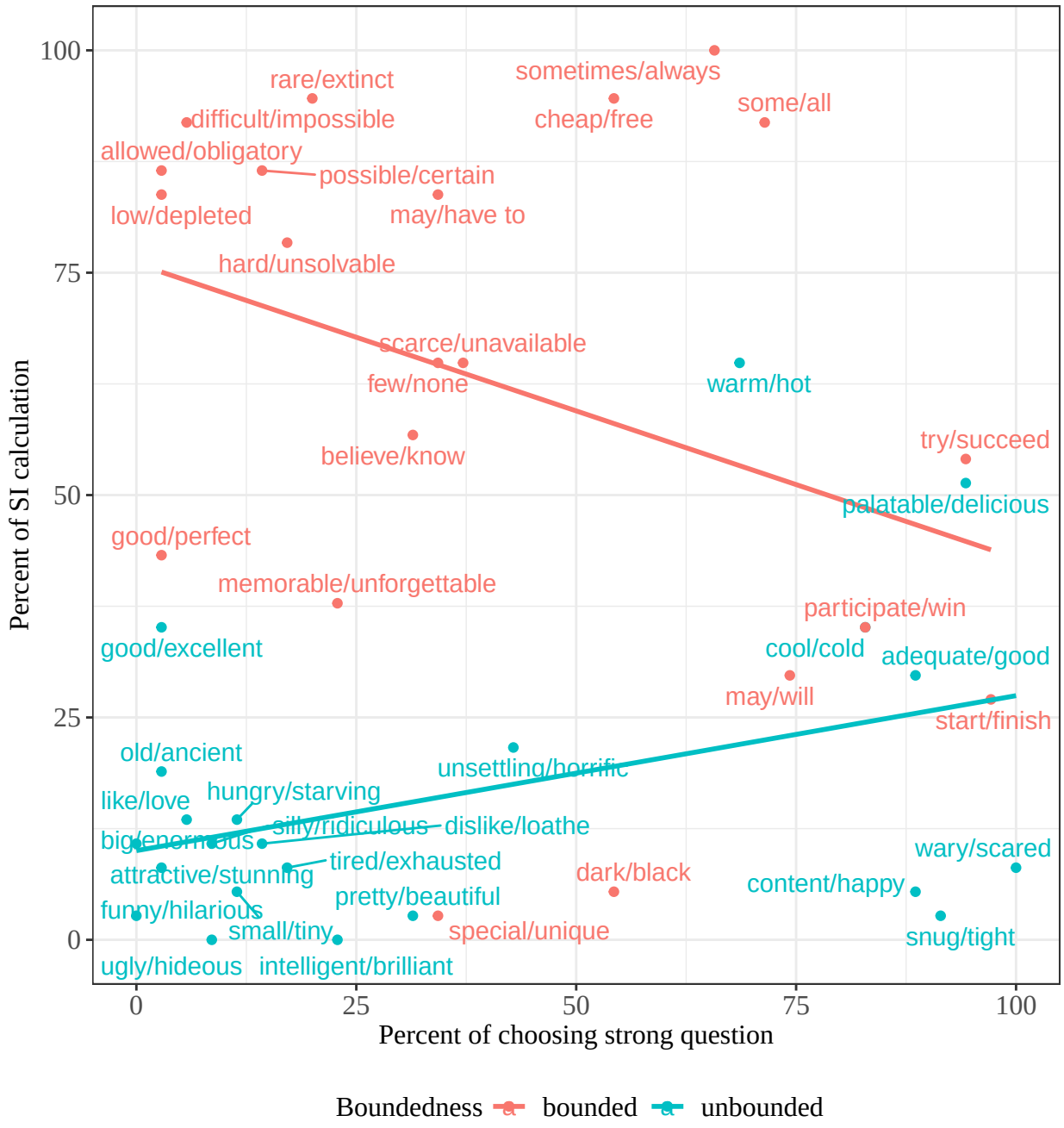

Figure 6. Results of Experiment 1 and 3. The x axis shows Question Availability from Experiment 3. The y axis shows SI Rate from Experiment 1.

5. General discussion. In this paper we tested the hypothesis that scalar diversity, in the absence of explicit discourse context, arises (in part) due to differences in what QUDs each potentially SI-triggering sentence brings to mind. On the one hand, this hypothesis predicts that once we provided explicit questions, that is, we made both strong- and weak-scalar questions 'equally available', scalar diversity would be eliminated. Though Experiment 2 provided evidence that an explicit QUD robustly influences SI calculation rates for a large number of scales, we still found robust variation in SI rates under both strong- and weak-scalar questions —an explicit QUD itself did not remove the scalar diversity effect.

On the other hand, in Experiment 3, we tested whether the differential availability of the two different polar questions can itself explain variability of SI rates. Though there was no general effect of Question Availability on SI rates, we did find that this metric predicted scalar diversity for a subset of scales. In particular, there was an interaction of Question Availability and boundedness: for unbounded scales, the more likely participants were to choose the strong-scalar question, the higher the corresponding SI rate was; in bounded scales, Question Availability had no effect. The exact nature of this interaction still needs further research, but we provide a possible explanation below.

In bounded scales, the stronger scalar term is not vague, but instead denotes a fixed point —more precisely, an endpoint. Thus this stronger scalar term is very salient as an alternative to the vague, weaker term (see e.g. van Tiel et al. 2016's distinctness). This high level of salience for the stronger alternative leads to high rates of SI calculation across the board, and Question Availability makes no difference for bounded scales. On the other hand, in unbounded scales, both scalar terms are vague: they denote intervals whose values vary according to context. The stronger scalar term on an unbounded scale is thus less salient as an alternative, and so it can experience a boost from context—in our case, from the availability of the strong-scalar question. Specifically, for unbounded scales, the more available a QUD based on the stronger term is, the more likely comprehenders will be to reason about that term as the stronger alternative. In turn, they will be more likely to derive the relevant SI.

It is worth noting that many of the stronger scalar terms in the unbounded scales that we examined are *extreme adjectives*, for example *starving*, *excellent*, and *tiny* (Morzycki 2012). For this subset of the scales we tested, where the stronger scalar is also an extreme adjective, a parallel can be drawn between our argument regarding unbounded scales and Morzycki (2012)'s proposal about extreme adjectives. In addition to predicating a gradable property of the subject, these adjectives also require that the subject's degree of that property be "off the scale" (see the denotation in (47) on p. 584). For Morzycki (2012), this is cashed out as an entailment that the subject's degree of that property is not already in the context.

Morzycki's analysis of extreme adjectives can help explain why SI rates vary depending on the availability of the strong-scalar question for unbounded, but not for bounded scales. Since part of the meaning of an extreme adjective is that the set of degrees it makes reference to are not in the discourse context, this adjective is not generally available as a salient stronger alternative for the purposes of SI calculation. This explains why the overall SI rate is low for scales where the stronger term is an extreme adjective; see also Gotzner et al. (2018) and Beltrama & Xiang (2013) for the same empirical finding. It then follows that the strong-scalar question will generally not be salient; it makes reference to a set of degrees not in the context. However, there might still be variation (as in our data) across these scales in how available the strong-scalar question is, which may then modulate the otherwise low rates of SI calculation:

the more available a question based on the "off the scale" scalar term is, the higher the rates of SI calculated from the weaker term will be.

Lastly, Beltrama (in press) also makes the suggestion that extreme adjectives (*non-logically extreme* in his terminology) show an interaction with discourse context that endpoint-denoting adjectives do not. He observes that the felicity of emphatic exclusives (e.g. *just* in *just amazing*) depends on the preceding context for extreme adjectives like *amazing*, but not for endpoint-denoting adjectives like *perfect*. Future research should investigate further the relationship between whether an expression is endpoint-denoting and how large a role context can play.

5.1. OPEN QUESTIONS. A number of open questions remain. In Experiment 2, we found that an explicit polar question that contains the stronger scalar term results in higher SI calculation rates than one that contains the weaker term. An important empirical follow-up would be to replicate this findings using QUDs that set up biasing contexts without explicitly mentioning one of the scalar terms. At present, our findings are also compatible with participants drawing relevance implicatures, and not SI, in the strong-scalar question condition. Consider, for instance, the dialogue in (9).

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

Here, if a participant concludes that Mary does not believe the student to be brilliant, this inference could have arisen as an effect of the Relation maxim (Grice 1967), even if <*intelligent, brilliant*> did not form a scale. Because Mary, in response to Sue's question which explicitly mentions *brilliant*, chooses not to directly agree or disagree, but rather to offer an alternative (*intelligent*), the negation of *brilliant* can be inferred irrespective of SI. Future work using different types of QUDs could distinguish this type of relevance implicature from genuine SI.

Future work should also test different empirical measures of Question Availability —the forced choice task employed in Experiment 3 was merely a first step in identifying such a measure. Additionally, it must also be kept in mind that Question Availability may itself be context-dependent; what QUD a conversational participant is most likely to entertain for a given, potentially SI-triggering, utterance may vary from context to context.

In Section 5, we began to explore the idea that boundedness and extremeness might both be relevant in explaining our finding that Question Availability predicts scalar diversity for only a subset of scales. However, our experimental items did include scales where boundedness and extremeness do not make the same predictions. For example, <*adequate, good*>, <*warm-hot*> and <*content-happy*> are all unbounded scales, but they are not extreme. Experiments that contain a larger item set and directly manipulate whether the stronger member of a scale is endpoint-denoting, extreme, or both, could shed more light on what subset of scales allow for a (larger) role of context in SI calculation. An additional consideration is whether extremeness as a notion can even be successfully applied to non-adjectival, e.g. verbal, modal, or quantifier, scales (though cf. Portner & Rubinstein 2016 for the argument that deontic modals can be extreme).

Finally, when taking all our findings together, an additional puzzle arises: while in Experiment 3 Question Availability showed a significant interaction with Boundedness, such that it was only a predictor of SI rates for unbounded scales, in Experiment 2 we found that context had an effect on (almost) all scales, irrespective of boundedness. Follow-up experiments

to both the QUD manipulation and to the empirical measure of Question Availability will shed light on whether these two findings can be reconciled.

6. Conclusion. In this study we investigated the role of QUD in explaining scalar diversity. Specifically, we tested the hypothesis that variation in SI rates, in the absence of explicit discourse context, depends on what kind of QUD the potentially SI-triggering sentences most naturally bring to mind. We first found that manipulating explicit questions (based on the stronger vs. weaker scalar) influences SI calculation rates for a large number of different scales. At the same time, however, there still remained substantial variation in SI rates across scales; providing an explicit QUD that contains the stronger scalar term did not lead to uniform SI rates. Nonetheless, the likelihood of asking a question based on the stronger scalar (Question Availability) was indeed found to be a factor contributing to scalar diversity, albeit only for unbounded scales. Taken together, our findings suggests that though an explicit QUD does not, in itself, eliminate the scalar diversity effect, contextual factors can explain some of the observed variability in how likely different scales are to lead to SI.

References

Beltrama, Andrea. In press. *Just* perfect, *simply* the best: An analysis of emphatic exclusion. *Linguistics and Philosophy*.

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

Cummins, Chris & Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Front Psychol* 6. 1779. https://doi.org/10.3389/fpsyg.2015.01779.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710. https://doi.org/10.1111/cogs.12171.

Doran, Ryan, Rachel Baker, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(2). 211–248. https://doi.org/10.1163/187730909x12538045489854.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb & Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88(1). 124–154. https://doi.org/10.1353/lan.2012.0008.

Drummond, Alex. 2007. Ibex Farm. http://spellout.net/ibexfarm.

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

Grice, Herbert Paul. 1967. Logic and Conversation. In Paul Grice (ed.), *Studies in the way of words*, 41–58. Cambridge, MA: Harvard University Press.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles: University of California dissertation.

Hulsey, Sarah, Valentine Hacquard, Danny Fox & Andrea Gualmini. 2004. The question-answer requirement and scope assignment. In Aniko Csirmaz, Andrea Gualmini & Andrew Nevins (eds.), *MIT Working Papers in Linguistics*, 71–90. Cambridge, MA: MITWPL.

Kuppevelt, Jan van. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19(4). 393–443. https://doi.org/10.1007/BF00630897.

Morzycki, Marcin. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natl Lang Linguist Theory* 30(2). 567–609. https://doi.org/10.1007/s11049-011-9162-0.
Portner, Paul & Aynat Rubinstein. 2016. Extreme and non-extreme deontic modals. In Nate Charlow & Matthew Chrisman (eds.), *Deontic modals*, 256–282. Oxford and New York: Oxford University Press.
Ronai, Eszter & Ming Xiang. 2020. Pragmatic inferences are QUD-sensitive: An experimental study. *Journal of Linguistics* 1st view. 1–30. https://doi.org/10.1017/S0022226720000389.
Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Front Psychol* 9. https://doi.org/10.3389/fpsyg.2018.02092.
van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. https://doi.org/10.1093/jos/ffu017.
Yang, Xiao, Utako Minai & Robert Fiorentino. 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. *Front Psychol* 9. 1720. https://doi.org/10.3389/fpsyg.2018.01720.
Zondervan, Arjen, Luisa Meroni & Andrea Gualmini. 2008. Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In Tova Friedman & Satoshi Ito (eds.), *Proceedings of Semantics and Linguistic Theory (SALT) 18*, 765–777. https://doi.org/10.3765/salt.v18i0.2486.