THE UNIVERSITY OF CHICAGO

SCALES, ALTERNATIVES, CONTEXT: EXPERIMENTAL INVESTIGATIONS INTO SCALAR INFERENCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE HUMANITIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF LINGUISTICS

BY
ESZTER RONAI

CHICAGO, ILLINOIS
AUGUST 2022

“Ez . . . önkényes és értelmetlen - de nem győzöm hangoztatni, hogy a nyelv döntően önkényes és értelmetlen (mivelhogy kommunikációs kód szegényke és nem a világ tükre és nem nemzetünk oltára).”

[“This is . . . arbitrary and senseless - but I cannot emphasize enough that language is overwhelmingly arbitrary and senseless (since the poor thing is a code for communication, and not a mirror of the world and not an altar to our nation).”]

–Ádám Nádasdy

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES ix

ACKNOWLEDGMENTS x

ABSTRACT xiii

1 INTRODUCTION 1
1.1 What can be an alternative 1
1.2 The role of context 2
1.3 Questions addressed in this dissertation 4

2 THE ACTIVATION OF SCALAR ALTERNATIVES IN PROCESSING 7
2.1 Alternatives in language processing 9
2.1.1 Focus alternatives 10
2.1.2 Scalar alternatives 13
2.2 Experiment 1: replication of Thomas et al. (2012) 15
2.2.1 Participants and task 15
2.2.2 Hypothesis and predictions 17
2.2.3 Results and discussion 17
2.3 Experiment 2: lexical priming 18
2.3.1 Participants and task 19
2.3.2 Hypothesis and predictions 19
2.3.3 Results and discussion 20
2.4 Experiment 3: sentential priming 21
2.4.1 Participants and task 21
2.4.2 Hypothesis and predictions 22
2.4.3 Results and discussion 22
2.5 Experiment 4: sentential priming with *only* 23
2.5.1 Participants and task 24
2.5.2 Hypothesis and predictions 24
2.5.3 Results and discussion 24
2.6 Summary of findings 25
2.6.1 Relevance for theory? 27
2.6.2 Remaining empirical puzzles 29
2.6.3 Future work 32

3 PROCESSING COST: CONTEXT OR ALTERNATIVE RETRIEVAL? 33
3.1 Background 34
3.1.1 The cost of implicature processing 34
3.1.2 Context effects on SI calculation 36
3.1.3 Processing cost: lexical access or context? 38

3.2 Contributions of Chapter 3 . . . 40
3.3 Experiment 5: QUD elicitation . . . 42
3.3.1 Participants . . . 42
3.3.2 Task, materials and procedure . . . 43
3.3.3 Results and discussion . . . 45
3.4 Experiment 6: QUDs modulate calculation rates and processing cost . . . 47
3.4.1 Participants . . . 47
3.4.2 Task, materials and procedure . . . 48
3.4.3 Predictions . . . 50
3.4.4 Results and analysis: rate of inference calculation . . . 51
3.4.5 Results and analysis: reaction times . . . 55
3.5 General discussion . . . 59
3.6 Summary of findings . . . 65

4 FACTORS EXPLAINING SCALAR DIVERSITY . . . 66
4.1 Scalar diversity . . . 67
4.1.1 Early evidence against uniformity . . . 68
4.1.2 Previous studies on scalar diversity . . . 70
4.2 New set of scales . . . 75
4.3 Experiment 7: Replicating scalar diversity . . . 77
4.3.1 Quantifying scalar diversity . . . 80
4.4 Explaining scalar diversity . . . 83
4.4.1 Experiment 8: Accessibility of stronger alternative . . . 83
4.4.2 Experiment 9: Distinctness of scalar terms . . . 87
4.4.3 Boundedness . . . 93
4.4.4 Frequency . . . 94
4.4.5 Semantic similarity . . . 96
4.4.6 Experiment 10: Meaning of the negated strong scalar . . . 100
4.4.7 Combined analysis and variance explained . . . 104
4.5 Summary of findings . . . 107

5 SEMANTIC AND PRAGMATIC EFFECTS ON SCALAR DIVERSITY . . . 109
5.1 Experiment 11: QUD elicitation . . . 109
5.1.1 Participants and task . . . 110
5.1.2 Results and discussion . . . 111
5.2 Experiment 12: Discourse context manipulation . . . 114
5.2.1 Participants and task . . . 115
5.2.2 Hypothesis and predictions . . . 116
5.2.3 Results and discussion . . . 117
5.3 Experiment 13: Focus manipulation . . . 121
5.3.1 Participants and task . . . 121
5.3.2 Hypothesis and predictions . . . 122
5.3.3 Results and discussion . . . 123
5.4 Experiment 14: Manipulating both cues . . . 130
5.4.1 Participants and task . . . 130

5.4.2 Hypothesis and predictions . . . 131
5.4.3 Results and discussion . . . 131
5.5 Summary of findings . . . 133

6 CONCLUSION . . . 134

REFERENCES . . . 136

LIST OF FIGURES

2.1 Results of Experiment 1: replication of Thomas et al. (2012) 18
2.2 Results of Experiment 2: lexical priming experiment testing scalar alternatives 20
2.3 Results of Experiment 3: sentential priming experiment testing scalar alternatives . . . 22
2.4 Results of Experiment 4: sentential priming experiment testing focus alternatives . . . 25
2.5 Results of Experiment 1-4 26

3.1 Example experimental trial from Experiment 5: SI (left) and EXH (right) 44
3.2 Example experimental trials from Experiment 6: SI (left) and EXH (right) 48
3.3 Results of Experiment 6: Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *scalar inference*. Different colors denote the different QUD conditions. Error bars represent standard error. 53
3.4 Results of Experiment 6: Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *it-cleft exhaustivity*. Different colors denote the different QUD conditions. Error bars represent standard error. 54
3.5 Results of Experiment 6: Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *scalar inference*. Different QUD conditions are displayed on separate plots. Colors denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error. 55
3.6 Results of Experiment 6: Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *it-cleft exhaustivity*. Different QUD conditions are displayed on separate plots. Colors denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error. 56
3.7 Example experimental trials from Experiment 6: SI (left) and EXH (right) 58

4.1 Example experimental trial from Experiment 7, which measured SI rates in a verification task. 78
4.2 Results of Experiment 7: Inference rates for 60 different scales. 79
4.3 Hypothetical SI rates: benchmarks for quantifying diversity using relative entropy. . . 82
4.4 Example experimental trial from Experiment 8 85
4.5 Results of Experiments 7 and 8: The *x* axis shows alternative accessibility from Experiment 8. The *y* axis shows SI rate from Experiment 7. 86
4.6 Example experimental trial from Experiment 9: weaker scalar term 89
4.7 Example experimental trial from Experiment 9: stronger scalar term 89
4.8 Results of Experiment 9: Average response on the degree estimate task (0-100 scale) for the two different conditions 91
4.9 Results of Experiments 7 and 9: The *x* axis shows distinctness between each weak-strong scalar pair from Experiment 9. The *y* axis shows SI rate from Experiment 7. . . 92
4.10 Results of the boundedness analysis: The *x* axis compares bounded and unbounded scales. The *y* axis shows SI rate from Experiment 7. 94
4.11 Results of the frequency analysis: The *x* axis shows the log-transformed frequency of the stronger scalar term. The *y* axis shows SI rate from Experiment 7. 95
4.12 Results of the semantic similarity analysis: The *x* axis shows the LSA similarity scores. The *y* axis shows SI rate from Experiment 7. 98

4.13 Example experimental trial from Experiment 10: stronger scalar term 101
4.14 Results of Experiments 9 and 10: Average response on the degree estimate task (0-100 scale) for the two different conditions 102
4.15 Results of Experiments 7 and 10: The *x* axis shows the meaning of the negated stronger term from Experiment 10. The *y* axis shows SI rate from Experiment 7. 103

5.1 Example experimental trial from Experiment 11: neutral condition 111
5.2 Example experimental trial from Experiment 11: SI-biasing condition 111
5.3 Example experimental trial from Experiment 12 115
5.4 SI rate for 60 different scales. Experiments 7 (SI) and 12 (Weak QUD, Strong QUD) are shown on the three facets of the plot. 118
5.5 Example experimental trial from Experiment 13 122
5.6 SI rate for 60 different scales. Experiments 7 (SI), 12 (Weak QUD, Strong QUD) and 13 (Only) are shown on the four facets of the plot. 124
5.7 Scalar analysis of the rank-order reading of *only*, adapted from Coppock and Beaver 2013, ex. 29 126
5.8 Scalar analysis of the complement-exclusion reading of *only*, adapted from Coppock and Beaver 2013, ex. 27 127
5.9 Example experimental trial from Experiment 14 130
5.10 SI rate for 60 different scales. Experiments 7 (SI), 12 (Weak QUD, Strong QUD), 13 (Only) and 14 (QUD + only) are shown shown on the five facets of the plot. 132

LIST OF TABLES

2.1 Average semantic similarity between prime-target pairs in the related and unrelated conditions in Experiments 1 and 2. 30

3.1 Predictions about the reaction time cost of implicature calculation 41
3.2 Results of Experiment 5: Frequencies of elicited question types 46
3.3 Results of Experiment 6: *Scalar inference*: Parameter estimates, standard errors, z values and p values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD. 52
3.4 *Results of Experiment 6: It-cleft exhaustivity*: Parameter estimates, standard errors, z values and p values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD. 52

4.1 Scalar expressions used in a representative sample of experiments on the interpretation, development and processing of scalar inferences. Table 2 from van Tiel et al. (2016, p. 139). 67
4.2 60 lexical scales collected. 76
4.3 Relative entropy results from Experiment 7 and the four hypothetical benchmarks . . . 83
4.4 Parameter estimates, standard errors, z values and p values from a logistic mixed effects regression model of the "Yes" vs. "No" responses in Experiment 7, predicted by the factors identified in this chapter 105
4.5 Parameter estimates, standard errors, z values and p values from a logistic mixed effects regression model of the "Yes" vs. "No" responses in Experiment 7, predicted by the factors identified in this chapter 106

5.1 Experiment 11 results: frequencies of elicited question types 113
5.2 Relative entropy results by experiment: Experiments 7, 12 and 13 123
5.3 Relative entropy results by experiment: Experiments 7, 12, 13 and 14 133

ACKNOWLEDGMENTS

Though various fragments of these acknowledgments have existed in my head for years, now that I have to make a coherent whole out of them, I find myself struggling. In this sense, the acknowledgements are quite similar to the actual dissertation itself.

Needless to say, the biggest thanks goes to Ming Xiang, who has been my advisor essentially from Day 1, and who has inspired and improved everything—from two qualifying papers, through three journal submissions (not counting the failed ones) and multiple dozens of experiments (including the failed ones), to this dissertation. Ming is famous for her unrelenting dedication to her students' projects, literally spending hours poring over data to solve some puzzle, or looking for a way to rephrase a complicated-sounding paragraph. But what I would like to highlight is Ming's incredible insight and sharpness. A quintessential feeling throughout my graduate school career has been that moment of realization —occurring quite often after several hours (or days) of frustration with some element of the research process —that something she has said was simply right. It has been a great privilege learning from you, Ming. Thank you for everything!

I am also indebted to my other committee members, Chris Kennedy and Itamar Francez. I am grateful for Chris's big-picture perspective: the way he has always managed to give me useful direction about framing, about finding an audience, about knowing what details I can afford not to worry about, and about recognizing when I really can't see the forest for the trees. I am not sure I can actually manage to do these things now, but I will always remember to try. I would like to thank Itamar for trying to ground me in linguistic theory, for bringing his healthy skepticism, and for always saying that a good dissertation is a finished dissertation. I will also never forget that when I panicked about missing semantics homework in my first year due to a bedbug infestation, he responded by suggesting that I could stay with him and his family.

Though he was not officially a member of any of my committees, my graduate school experience and research have been shaped in innumerable ways by Karlos Arregi. Karlos has generously advised my forays into syntax, bringing his incredible insight and patience to various projects. To this day, I'm sad that Karlos doesn't do experimental pragmatics, and I wonder whether I should have attempted a syntax dissertation just so I could have asked him to be on my committee. I would also like to thank Alan Yu for all he has taught me about experimental methods and professionalization, as well as Salikoko Mufwene and Allyson Ettinger for their guidance as faculty sponsors of the Language Evolution Acquisition and Processing workshop.

I am very lucky to have benefited from interactions with and feedback from linguists from other institutions—at workshops, conferences, and through their visits to UChicago. I want to thank Hannah Rohde, Yi-Ting Huang, Bob van Tiel, Brian Dillon, Jeffrey Witzel, Masaya Yoshida, John Tomlinson, Laura Staum Casasanto, Masha Polinsky, Dan Yurovsky, Tim Leffel, Marcin Morzycki, and Jinghua Ou, as well as the organizers and mentors of the Pop-Up Mentoring Program of the Committee on Gender Equity in Linguistics. The last few months of my dissertation writing were spent in Paris, as a visitor at the Laboratoire de Linguistique Formelle; I am grateful to Barbara Hemforth and Ira Noveck (and his GRISP group) for making that possible and welcoming me there. Lastly, I would not have ended up in graduate school in the first place without the support and encouragement of my undergraduate teachers in Cambridge: Kasia Jaszczolt, Theresa Biberauer, Moreno Mitrović, and Chi-Hé Elder.

The real treasure of grad school is certainly the friends I made along the way. In particular, I couldn’t have asked for better cohortmates than Laura Stigliano and Yenan Sun. They have been wonderful friends, conference (organizer) buddies, gossip aficionados, chat group namers, and, last but not least, co-authors. Here’s to many more years of singing Material Girl together! I am also thankful for the friendship and support of Daniel Lam (who got me into organizing LEAP, into bouldering, and into many more exciting things I now forget), Sanghee Kim (who completes our Ming’s Minions group) and honorary cohortmate Anisia Popescu—as well as Emre Hakgüder, Aurora Martinez del Rio, Matt Hewett, Jackie Lai, Amara Sankhagowit, and Thomas Sostarics. I would like to thank for their friendship and mentorship the more senior (now former) graduate students at UChicago: Andrea Beltrama, Jeff Geiger, Helena Aparicio, Jessica Kantarovich, Ksenia Ershova, Julian Grove —and especially Adam Singerman. For me, one of the most enjoyable aspects of academia is having a larger cohort of friends and colleagues to continually run into and share research, gossip and Ethiopian food with; I am lucky to have met—and hope to continue meeting—Matt Tyler, Aniello de Santo, Alex Göbel, Sherry Yong Chen, Lelia Glass, Martín Fuchs, Josh Phillips, Jonathan Pesetsky, and Marju Kaps.

This list of acknowledgements would not be complete without my non-linguist friends, who have provided me with support long before I even knew linguistics existed, and many of whom even made it to my dissertation defense: Zsolt Verasztó (the secret co-author on more of my projects than I will ever admit), Eszter Lévai, Réka Mándoki, Rozi Vőfély, Bence Börcsök, Balázs Mezei, Balázs Dezsényi, Dani Szetei, Máté Krausz, Juci Pétervári, and Lea Bálint. I am also grateful for the friendship of my undergraduate linguistics friends, the lingling tingtingz, especially Meg Proops, as well as honorary linguist Amrita Dasgupta.

Last but not least, I would like to thank my family, as well as my family-in-Chicago, the Tabatowskis. I am especially thankful for the support of my brother—who undoubtedly launched me on this career path by telling me (when I was 14) to ignore everything other than math and English, and then by noting (when I was 17) that someone who writes about semantic transparency shouldn't become an economist; my sister, who has always been there for me for everything non-work-related; my dad, who enjoys counting up the number of PhDs at family events and has long anticipated mine; and my mom, who (not easily, but eventually) allowed me to not stay at home hogy a kisszobában varrogassak. Mike Tabatowski deserves his own paragraph, but I am not a good enough writer to give him one. (Perhaps that's why I make him write at least 60% of all text I produce.) Thank you for joining me not only on the journey of graduate school, but the much more important journey of life, and for always being there as my silent partner and my real partner; I truly couldn't have done it without you.

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

ABSTRACT

Scalar inference, the process by which we infer meanings stronger than what was explicitly said, has long been a central topic of investigation in theoretical semantics-pragmatics, as well as in psycholinguistics. Upon encountering the sentence *Mary ate some of the deep dish*, for instance, hearers regularly compute the pragmatic meaning that Mary ate some, but not all, of the deep dish. The standard assumption is that the inferential process that gives rise to this result involves hearers reasoning about what the speaker could have said, but did not (Grice, 1967). Further, Neo-Gricean accounts typically assume that hearers infer the negation of informationally stronger unsaid alternatives, e.g., because <*some, all*> form a lexical scale, and *all* is stronger than *some*, hearers derive *not all* upon encountering *some* (Horn, 1972; Katzir, 2007). In addition to involving the negation of unsaid alternatives, another crucial property of scalar inference is context-sensitivity (Van Kuppevelt, 1996). That is, whether a scalar inference-enriched meaning is derived depends partially on the discourse context. For instance, given a question such as *Did Mary eat any of the deep dish?*, all that matters is whether Mary ate at least some of the deep dish. Therefore, in this context, an answer of *She ate some of the deep dish* is less likely to lead to scalar inference, since *She ate all of the deep dish* is no longer a relevant alternative.

This dissertation is an experimental investigation into these two crucial properties of scalar inference: alternative- and context-sensitivity. In Chapter 2, I test whether alternatives such as *all*, which are important in the theoretical modelling of scalar inference, are psychologically real. In a series of semantic priming with lexical decision experiments, I demonstrate that such lexical alternatives are indeed retrieved and activated in the processing of inference-triggering utterances. In Chapter 3, I turn to the much-investigated question of whether scalar inference calculation incurs a processing cost, operationalized as increased reaction times. In a sentence-picture verification experiment, I compare scalar inference with another pragmatic inference, *it*-cleft exhaustivity, and demonstrate that whether inference calculation is costly is a function of the discourse context, and not whether alternative-retrieval is involved.

While much research has investigated the *some but not all* inference, there exist many other scales where a set of lexical items are ordered with respect to each other in terms of their logical strength. Similarly to <*some, all*>, <*good, excellent*> also form a scale; an utterance of *The movie is good* might give rise to the scalar inference that the movie is not excellent. An important recent finding, however, is that there is in fact considerable variation across such different scales in the likelihood of inference calculation: the *not excellent* inference, for instance, is much less likely to arise than *not all* (van Tiel et al., 2016). The second half of this dissertation investigates this phenomenon of *scalar diversity*. In Chapter 4, I ask whether the observed variation can be explained by differences in the alternatives themselves (*all* vs. *excellent*). My findings suggest that some (but definitely not all) properties of alternatives (e.g., accessibility) do indeed predict likelihood of inference calculation, yet scalar diversity remains largely unexplained. Lastly, in Chapter 5, I investigate whether scalar diversity can be modulated by a supportive discourse context vs. grammatically encoding the negation of alternatives, and show that uniformity in inference calculation is only achieved when these two factors align.

Overall, the picture that emerges from this dissertation is that pragmatic meaning arises both as a function of global properties of context, and as a function of local properties of the scalar terms themselves.

CHAPTER 1

INTRODUCTION

In natural language communication, inferred messages do not always equal literal messages. One well-studied instantiation of this phenomenon is (the family) of implicatures, exemplified in (1) by scalar inference (SI).

(1) Mary ate some of the deep dish.
Literal meaning: Mary ate at least some of the deep dish.
Inference-enriched meaning: Mary ate some, but not all, of the deep dish.

It is commonly assumed that SI arises because hearers consider and reason about informationally stronger alternatives that could have been uttered in place of what has actually been uttered. In particular, upon encountering the utterance in (1), hearers are taken to reason that the stronger statement *Mary ate all of the deep dish* was also available to the speaker. Because the speaker chose not to utter this stronger alternative, the hearer can infer as an SI its negation (*Mary didn't eat all of the deep dish*), enriching the literal meaning of (1) to its inference-enriched meaning. This process can be viewed as an interaction of the Quality and Quantity maxims (Grice, 1967). Speakers make contributions that are as informative as required by their dialogue context (but not more so; Quantity), but they will not say what they believe to be false or what they lack evidence for (Quality). *Mary ate all of the deep dish* is more informative than what was actually uttered, and therefore the speaker should have uttered this stronger alternative in accordance with Quantity. But she did not, which suggests to the hearer that she did not believe it to be true (Quality).

1.1 What can be an alternative

There are various limits on when SIs actually arise, or what can serve as a stronger alternative. For instance, SI is predicted not to arise when the speaker is not knowledgeable about the informationally stronger alternative. In this case, uttering the alternative would have been a violation of the

Quality maxim ("Do not say that for which you lack adequate evidence."), and not uttering it is not a signal of its falsity. Speakers are also taken to obey maxims other than Quality and Quantity, e.g., they make their utterances relevant (Relation). SIs are therefore also predicted not to arise when the informationally stronger alternative would not have been relevant in the present discourse context. In that case, if the speaker were to utter the stronger alternative, she would violate Relation, and therefore not uttering it is, again, not meant to signal its falsity.

Narrowing down "what could have been uttered", that is, establishing what constrains the possible unsaid alternatives that hearers reason about, has in large part been motivated by the so-called symmetry problem (Kroch, 1972; Horn, 2000a; von Fintel and Fox, 2002; Katzir, 2007; Fox and Katzir, 2011). The symmetry problem is exemplified in (2):

(2) a. Mary ate some of the deep dish.
b. Mary ate all of the deep dish.
c. Mary ate some but not all of the deep dish.

As we have seen, when calculating the inference-enriched meaning of (2-a), hearers reason about the alternative in (2-b) and infer its negation. In principle, though, (2-c) could also serve as an alternative to (2-a). However, if hearers reasoned with (2-c) as an alternative to (2-a), they should arrive at the enriched interpretation *Mary ate all of the deep dish*, while the same reasoning applied to (2-b) derives the familiar *Mary ate some but not all of the deep dish* interpretation. The two enriched interpretations are mutually exclusive. Theories of what constitutes an alternative thus aim, at minimum, to rule out (2-c) as a possible alternative.

1.2 The role of context

Another important property of SI is context-sensitivity (Van Kuppevelt, 1996). One way to formalize this is by reference to the concept of Question Under Discussion (QUD), defined as the immediate topic of discussion that proffers a set of relevant alternatives (Roberts, 1996/2012). To illustrate how this would work for our original example from (1), consider the following dialogue:

(3) A: Did Mary eat all of the deep dish?
B: She ate some of the deep dish.

Standard semantic treatments of questions take them to partition a set of possible worlds into cells denoting their possible answers (Hamblin, 1976; Groenendijk and Stokhof, 1984). The question *Did Mary eat all of the deep dish?*, then, partitions the Common Ground based on the stronger alternative *all*: in one cell are all the worlds where Mary ate all of the deep dish (4-a), and in the other cell, all the worlds where she did not eat all of the deep dish (4-b).

(4) a. $\{w : \forall x.x$ is a portion of the deep dish $\rightarrow$ Mary ate x in $w\}$
(*Mary ate all of the deep dish.*)
b. $\{w : \neg\forall x.x$ is a portion of the deep dish $\rightarrow$ Mary ate x in $w\}$
(*Mary did not eat all of the deep dish.*)

An answer, in turn, is taken to be congruent with (or “a good answer to”) a question if it determines which cell contains the actual world (Hulsey et al., 2004). Consider, then, the two readings (literal and inference-enriched) of the sentence *Mary ate some of the deep dish*. Given A’s question in (3), the inference-enriched meaning (*Mary ate some, but not all, of the deep dish.*) is a congruent answer, because it entails the “not all” cell of the partition (4-b), and eliminates the “all” cell (4-a). The literal meaning (*Mary ate at least some of the deep dish.*), on the other hand, does not entail either cell, and it therefore does not directly bear on the question. Therefore, only on its inference-enriched meaning does the sentence constitute a congruent answer, and a discourse context such as (3) encourages SI calculation.

The picture is different, however, given a dialogue like (5)

(5) A: Did Mary eat any of the deep dish?
B: She ate some of the deep dish.

A polar question like *Did Mary eat any of the deep dish?* partitions the Common Ground such that in one cell are all the worlds where Mary ate some of the deep dish (6-a), and in the other cell, all the worlds where Mary did not eat any of the deep dish (6-b).

(6) a. $\{w : \exists x.x$ is a portion of the deep dish and Mary ate x in $w\}$
(*Mary ate some of the deep dish.*)

b. $\{w : \neg\exists x.x$ is a portion of the deep dish and Mary ate x in $w\}$
(*Mary did not eat any of the deep dish.*)

Here, both readings (literal and inference-enriched) of the potentially SI-triggering sentence (*Mary ate some of the deep dish*) constitute a good answer, since both entail the (6-a) cell of the partition. Therefore, a discourse context such as (5) does not particularly encourage SI calculation.

We can see, therefore, how the same sentence can be more or less likely to give rise to SI depending on the discourse context.

1.3 Questions addressed in this dissertation

This dissertation will investigate a range of questions related to the above two important properties of SI: alternative- and context-sensitivity. The overarching goal of the dissertation is to assess the respective roles and importance of alternatives and discourse context in the calculation and psycholinguistic processing of SI.

Chapter 2 investigates the psycholinguistic mechanisms underlying SI. Specifically, it tests whether lexical alternatives (like *all*) are retrieved and activated when hearers process sentences that (potentially) trigger SI calculation. In a series of semantic priming with lexical decision experiments, I find that alternatives like *all* are indeed activated in the real-time processing of SI, suggesting that lexical scales are psychologically real. These findings show that scalar alternatives are processed similarly to focus alternatives, whose exclusion is grammatically, rather than pragmatically, encoded.

Chapter 3 addresses the much-investigated question of whether SI calculation incurs processing cost, or on the other hand, whether SI is a cost-free default inference. In this chapter, I compare SI to another pragmatic inference, *it*-cleft exhaustivity. Importantly, the calculation of *it*-cleft exhaustivity is taken to proceed without reference to lexical alternatives. Therefore, if processing cost arose as a result of having to retrieve and reason about alternatives, only SI should be costly, and *it*-cleft exhaustivity should be cost-free. Instead, the findings of my sentence-picture verification experiment show that for both pragmatic inferences, the likelihood of inference calculation, as well as the associated reaction times, vary depending on the QUD. This suggests that the processing cost of pragmatic inference calculation is better explained by properties of the discourse context than by whether alternative-retrieval is involved.

In Chapter 4, I investigate a recent puzzle for SI, termed *scalar diversity*. Scalar diversity is the observation that lexical scales vary substantially in how likely they are to give rise to SI. For instance, the *some but not all* SI is much more likely to arise than the parallel inference *good but not excellent*. This is somewhat unexpected, since *excellent* is an informationally stronger alternative to *good* in the same way as *all* is an alternative to *some*. This chapter tests whether scalar diversity can be explained by properties of the alternatives themselves. In other words: can the differential likelihood of e.g., the *some but not all* vs. the *good but not excellent* SI be predicted by some difference between *all* and *excellent*? A number of experiments are conducted, testing relevant factors proposed in existing literature (some operationalized in novel ways), as well as new ones. Findings reveal that some properties of alternatives (accessibility of the alternative, distinctness of the weaker term and the alternative, boundedness of the scale) do indeed serve as predictors of scalar diversity, but some others (frequency, semantic similarity between the weaker term and the alternative) dot not. Overall, a lot of the statistical variance remains unexplained, and the best predictor in fact turns out to be one not directly related to properties of alternatives.

Finally, Chapter 5 turns to how alternatives and context can modulate scalar diversity. Specifically, I test whether the observed inter-scale variation in inference rates changes when inference-triggering sentences are placed under a biasing QUD (e.g., *Did Mary eat all of the deep dish?*) or include the focus particle *only* (*Mary ate only some of the deep dish*), which encodes alternative exclusion in the grammar. I show that both of these manipulations result in increased rates of inference calculation and reduced variation across scales —as quantified by the proposed information-theoretic measure of relative entropy. But under both the QUD and the focus manipulation, scalar diversity is not fully eliminated. However, when the two investigated factors align, i.e. there is pragmatic support from the context and a semantic cue to exclude alternatives, we do find uniformly high rates of inference calculation.

Taken together, the experiments reported in this dissertation suggest an important role for both discourse context and alternative exclusion in explaining SI calculation and processing, and variation therein.

CHAPTER 2

THE ACTIVATION OF SCALAR ALTERNATIVES IN PROCESSING

Chapter 2 investigates the psycholinguistic mechanisms underlying scalar inference[1]. The standard assumption is that the inferential process that gives rise to SI involves hearers reasoning about what the speaker could have said, but did not. But it is an open question precisely what psycholinguistic mechanisms underlie this inferential process. In this chapter I use semantic priming with lexical decision to test whether lexical alternatives are retrieved and activated in the processing of SI-triggering sentences. My findings suggest that comprehenders indeed activate the alternatives that theories of SI would take them to reason about; in other words, lexical scales are psychologically real.

Previous work that tested the activation of scalar alternatives using semantic priming has also related findings to different theoretical accounts of SI. In particular, such findings have been used to adjudicate between Neo-Gricean and Post-Gricean theories of SI. For this reason, it is worth briefly reviewing the differences between these two families of accounts.

Neo-Gricean accounts typically assume that hearers infer the negation of informationally stronger alternatives that the speaker could have said, and that these alternatives are determined via the lexicon or grammar —e.g., because <*some, all*> form a lexical scale, and *all* is stronger than *some*, hearers derive *not all* upon encountering *some*. In more detail, Horn (1972)'s influential proposal posits that some lexical items are specified as belonging to a scale, i.e. a set of lexical items which are ordered with respect to each other in terms of their logical strength. Horn's proposal formalizes the informativity requirement (that the alternative be "stronger") as asymmetric entailment: a sentence like *Mary ate all of the deep dish* entails *Mary ate some of the deep dish*, but not vice versa. Additionally, scales must be monotonic.

There exist other proposals that take as their starting point the existence of Horn-scales, and introduce further constraints on them. The intuition that SIs are not derived if they would conflict

[1] Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/wga25/?view_only=cefc447bc6e649aeb4815de958d71597

with another Gricean maxim (such as Relation) is spelled out i.a. by Matsumoto (1995)'s Conversational Condition on scales, which states that for an SI to be licensed, the choice of a stronger scalar term must not be attributed to the observance of any information-selecting maxim other than Quality and Quantity-1. Hirschberg (1985) observes that other than lexical items that asymmetrically entail one another, rank orderings, spatial orderings and process stages also form a scale. Gazdar (1977) posits that two terms must also share presuppositions in order to count as scalar alternatives. Atlas and Levinson (1981) propose that scalar alternatives must also belong to the same semantic field and be lexicalized to the same degree. At the core of all these proposals is the existence of lexical scales.

As mentioned, the other family of Neo-Gricean accounts is one proposing structurally-based alternatives. Katzir (2007) argues for the following structure-sensitive characterization of alternatives: for a structure ϕ, the alternatives are those structures which are at most as complex as ϕ. Specifically, this means that constituents of the uttered sentence can be deleted, contracted, or substituted from the lexicon (or with material from the original utterance) to derive alternative(s). In scalar inference, *Mary ate all of the deep dish* needs to be constructed as an alternative to *Mary ate some of the deep dish* in order to derive the familiar *some but not all* inference. Alternative construction in this case requires the operation of substituting *some* with another constituent from the lexicon, viz. *all*. This can be contrasted e.g. with free choice inference (where a disjunctive sentence is embedded under an existential modal operator), in which case the alternatives are constructed via the deletion of one of the disjuncts from the uttered sentence. This account thus contrasts with the Horn-scale-based accounts in that the possible set of alternatives is restricted based on the syntax, and not the lexicon.

Substantially different from Neo-Griceans are Post-Gricean accounts, such as Relevance Theory (Sperber and Wilson, 1995). According to Relevance Theory, the final interpretation of an utterance is obtained through a contextually driven process of ad hoc concept construction (loosening, or in the case of SI, strengthening), which is on par with loose talk or metaphors. As such, SI calculation is not a lexically based process; rather the construction of ad hoc concepts is at the core of it, with lexical scales having no special role. Such an account instead attributes great importance to context in determining the set of relevant alternatives.

This brief background of Neo- vs. Post-Gricean accounts will help interpret some of the findings of previous studies. In Section 2.6.1, after the presentation of my own experiments, I will further discuss what relevance priming results can have for theory. But first, this chapter starts with a broader introduction to previous work on the processing of alternatives (Section 2.1). I will then present the findings of four semantic priming experiments: Experiment 1 is a replication of semantic priming unrelated to SI (Section 2.2); Experiment 2 tests scalar alternatives in a semantic priming experiment without sentential context (Section 2.3); Experiment 3 adds sentential context and tests activation in potentially SI-triggering sentences (Section 2.4); finally, Experiment 4 tests the activation of alternatives in the presence of the focus particle *only*. The chapter concludes with the discussion of some remaining empirical puzzles (Section 2.6.2).

2.1 Alternatives in language processing

Alternatives are pervasive in (the modeling of) semantic-pragmatic meaning, including, but not limited to scalar inference. Correspondingly, alternatives have generated a lot of interest in psycholinguistics too, with various experimental paradigms being used to probe what kind of mental representations they have. For instance, a counterfactual statement like *If I lived in Chicago, I would eat deep dish pizza every day* evokes possible worlds where I live in Chicago (and suggests that these are non-factual worlds), as well as ones where I do not live in Chicago (which include the actual world). Similarly, if I say *I don't live in Chicago*, I am asserting that in the actual world, the proposition that I live in Chicago is false, but this statement also brings to mind worlds in which I actually do live in Chicago. In both constructions, then, different sets of alternative worlds are juxtaposed. Indeed, a lot of studies have investigated the processing of such alternative-evoking constructions, with particular attention to e.g. the timecourse of activating alternatives —among many others, see Kaup et al. (2007); Kaup and Zwaan (2003); Tian et al. (2016) on the processing of negation, and Ferguson et al. (2008); de Vega and Urrutia (2012) on counterfactuals.

The alternatives evoked in negation and counterfactuals are not constrained by the lexicon, while this dissertation focuses on lexical alternatives involved in scalar inference. Lexical alternatives are similarly relevant in (the processing of) sentential focus. Therefore, in the following sections, I discuss existing work in more detail on the processing of focus and scalar alternatives. For recent overviews on the role of alternatives in language processing, see Repp and Spalek (2021) and Gotzner and Romoli (2022) (and references therein).

2.1.1 Focus alternatives

Sentential focus, i.e. the marking of new or emphasized information in a sentence is a core linguistic device used to structure the flow of conversation (Chomsky, 1972; Jackendoff, 1974). This information is often provided in (implicit) contrast to possible other alternatives (Rooth, 1992, 1985). In English, contrastive focus can be marked by placing a prominent accent on a word:

(1) Mary ate DEEP DISH.

Focus is invoked also in the computation of the meaning of sentences that contain focus-sensitive operators (or focus particles), such as *only*:

(2) Mary only ate DEEP DISH.

Both (1) and (2) convey what Mary ate (deep dish), and crucially also that Mary did not eat anything else from among a set of contextually-determined of alternatives (e.g., {thin crust pizza, lasagne, salad, ...}). In successful comprehension, hearers infer this set of contrastive alternatives as intended by the speaker.

A number of psycholinguistic studies have shown that focused elements (*deep dish* above) are remembered more accurately and with more semantic detail than non-focused elements (Birch and Garnsey, 1995; Sturt et al., 2004). Importantly, it has also been found that memory representations are improved for focus alternatives as well. The logic behind experimental investigations of focus alternatives is that if alternatives are reactivated or retrieved during the comprehension and processing of some structure, then they should also be remembered and therefore recalled better. That is, focus-sensitive expressions are predicted to improve memory for both the focused element and for information-structural alternatives, as compared to a neutral baseline.

Fraundorf et al. (2010) tested intonational focus. The authors provided participants with a context story where a constituent had pitch accent, then probed recognition memory accuracy via forced choice and true/false judgement tasks. They found that pitch accent increased hits to correct statements and correct rejections of a contrast item, while unrelated items were unaffected by the experimental manipulation. This suggests that focus results in an enhanced representation of what happened and also what did not happen (but could have happened). In other words, comprehenders use pitch accenting to encode and update information about multiple elements in a contrast set: prosodically marked focus was found to modulate the representation of both the accented word itself and a contrasting alternative. In a follow-up study, Fraundorf et al. (2013) showed that similar results obtain if prominence is marked using font emphasis, rather than pitch accenting.

Spalek et al. (2014) tested focus particles, and conducted an experiment in the delayed recall paradigm. Auditory stimuli were used to first introduce a set of elements (e.g. peaches, cherries, bananas), and then continue with either the exclusive particle *nur* (*only*), the inclusive German particle *sogar* (*even*), or no particle (neutral control)—see (3) below.

(3) Speaker 1: In the fruit bowl, there are peaches, cherries, and bananas. I bet Carsten ate cherries and bananas.

Speaker 2: No, he only/even/∅ ate peaches.

After ten trials, participants had to recall the elements in the context sentence. The authors found that both particles enhanced memory performance for the focus alternatives (above: cherries, bananas), relative to the control condition. Note that the two particles differ in their meaning: *only* specifies an exclusive contrast such that a property of the referent does not hold of its alternatives, whereas *even* specifies an additive contrast, where the referent's property is shared with its alternatives. Crucially, however, the particles are similar in that they both make reference to alternatives. Spalek et al.'s findings thus suggest that information-structural alternatives are better encoded in memory when a focus-sensitive particle is present. See also work by Gotzner and Spalek (2017), which also tested focus particles, but used the probe recognition paradigm.

For further experimental paradigms, as well as ways of marking sentential focus, see i.a., Sanford et al. (2009) for findings about cleft sentences, using a change detection task, and Kim et al. (2015), who tested the English focus particles *only* and *also* in the visual world eye-tracking paradigm.

A number of studies testing the processing of sentential focus, and the activation of focus alternatives, have used semantic priming tasks. Husband and Ferreira (2015) (following Braun and Tagliapietra 2009; see also Gotzner et al. 2016; Yan and Calhoun 2019) used lexical decision with cross-modal priming to investigate focus alternatives. Participants were auditorily presented with sentences such as (4), and had to make a decision about whether a visually presented target word was an English word.

(4) The murderer killed the NURSE last Tuesday night.

The prime in each sentence was the focused element (*nurse* in (4)), while the targets in the lexical decision task were contrastive semantic associates (focus alternatives, e.g. *doctor*), non-contrastive semantic associates (e.g. *clinic*) and unrelated words. The study found early activation (i.e. facilitated reaction times in the lexical decision task) of both contrastive and non-contrastive semantic associates in sentences where *nurse* was focused. However, later activation (after a longer stimulus onset asynchrony) was only found for contrastive alternatives, suggesting that the initial activation of non-contrastive semantic associates decays.

My experiments on scalar alternatives that are reported in this chapter are modeled after the semantic priming experiments that tested focus alternatives. But before introducing those experiments, I briefly summarize existing work that has also used priming to investigate SI.

2.1.2 Scalar alternatives

In this section I review existing priming studies in the domain of scalar alternatives, in relation to my own experiments reported later in the chapter.

De Carvalho et al. (2016) tested 18 different lexical scales, and used lexical decision with masked (subliminal) priming to see if one member of a scale (e.g., *some*) activates the other (*all*). In their experiment, participants were visually presented with a prime word for a very short amount of time (34ms, too quick for the participant to realize it was presented), and then had to make a decision about whether the following visually-presented target word was a word of English. A priming effect should manifest as facilitated response times in making the lexical decision. This study used priming to adjudicate between different theories of SI. Specifically, the authors made the assumption that under a Neo-Gricean account of SI, which relies on lexically-given Horn-scales, the stronger alternative *all* is always needed in the processing of the weaker term *some*. On the contrary, *some* is not needed to process *all*. This makes the prediction that *some* would prime *all* more than *all* would prime *some*. A Post-Gricean account such as Relevance Theory, on the other hand, does not assign special significance to lexical scales. The authors therefore made the further prediction that under Post-Gricean accounts, any priming effect should merely be due to semantic relatedness and not show asymmetry, i.e., *some* would prime *all* only to the extent that *all* also primes *some*. The findings are in line with the Neo-Gricean account. An important difference between de Carvalho et al.'s work and my goal in this chapter (as well as the literature on focus alternatives discussed above), is that de Carvalho et al. tested whether scalar terms prime each other in the absence of any sentential context that would trigger SI calculation. In contrast, what my experiments will primarily be testing is whether scalar alternatives are primed in sentences that (might) trigger SI calculation.

Schwarz et al. (2016) also used subliminal priming (prime presentation for 32/48ms), and they tested 28 different (adjectival) scales. Their goal with priming was not to test whether a particular alternative is evoked in language processing, but rather to increase the salience/availability of scalar alternatives. Specifically, in their experiment, participants were primed with the stronger alternative before making an SI judgment: for instance, the sentence *The task is difficult* can lead to the SI that the task is not impossible. In one condition of Schwarz et al.'s experiment, participants saw the stronger alternative *impossible* as a prime before reading the sentence *The task is difficult* and indicating whether they calculated the *not impossible* SI from it. The particular hypothesis tested was that when the stronger alternative is made salient via priming, participants would be more likely to calculate the SI, and would do so faster. (This, in essence, is a novel operationalization of van Tiel et al.'s (2016) hypothesis of alternative availability, which will be discussed in detail in Section 4.1.2 of Chapter 4.) The strong alternative prime condition (*impossible)* was compared to conditions where the prime was an opposite (*easy*), or the word itself, i.e., identity priming (*difficult*), or a control (+++++++). Schwarz et al.'s findings were not in line with the hypothesis that priming with the stronger scalar alternative makes SI calculation likelier and faster, though there were other effects in their data that suggest that priming successfully influenced participants' responses in other ways.

Lastly, there is also a growing body of work that uses priming in investigations of SI, but not in the sense of testing whether there is semantic priming between scalar terms. Instead, these studies use so-called structural priming. For instance, participants have to make judgments on sentences that are true on their literal, but not their SI-enriched meanings, such as *Some sheep are mammals*. But before doing so, they are presented with a true sentence that contains the alternative *all*, such as *All lions are mammals*. It has been found that when participants have been primed with such a sentence, they are subsequently more likely, and faster, to calculate the SI from *Some sheep are mammals* (Rees and Bott, 2018; Bott and Frisson, 2022). Similarly, SI calculation can also be primed by pairing SI-triggering sentences with images that force participants to interpret those sentences as either *not all* or *all* (i.a., Bott and Chemla 2016). The effect of the structural priming paradigm can be interpreted as the stronger alternative (*all*) being primed, or the mechanism of SI calculation itself being primed. In the latter case, the idea is that once a participant has gone through the SI calculation process for a prime trial (e.g., one where the SI-enriched interpretation was forced via a picture), they will be more likely to do so on subsequent trials.

As we will see, the experiments I report on in this chapter are most similar in their design to the priming studies on focus alternatives (e.g., Husband and Ferreira 2015), rather than the priming studies on scalar alternatives just summarized —though it will be informative to compare my findings to de Carvalho et al. (2016), and as mentioned, the availability hypothesis tested by Schwarz et al. (2016) will become relevant in Chapter 4.

2.2 Experiment 1: replication of Thomas et al. (2012)

Given that priming experiments are typically conducted in person in a lab setting, and web-based priming is less commonly done, I first conducted a replication experiment on stimuli unrelated to scalar inference. This was to confirm that the web-based platform (and timing parameters) I adopt in later experiments can reliably reveal effects of semantic priming.

2.2.1 Participants and task

50 native speakers of American English participated in an online experiment, administered on the PCIbex platform (Zehr and 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 removed if their accuracy on the lexical decision task was below 90%. Data from 39 participants is reported below.

Experiment 1 was a semantic priming with lexical decision experiment. Participants had to decide whether a word they saw was a word of English or not; this word was the target. They indicated their decision by pressing the F key for "non-word" and the J key for "word". The primary dependant variable of interest is their reaction time in making this lexical decision. Participants were instructed to make a decision as fast as possible, while remaining accurate. Crucially, before making a lexical decision on the target, participants were presented with another word, the prime. There were two experimental conditions: in the "related" condition, the target word (e.g., *boy*) was preceded by a prime word that was semantically related to it (*girl*). In the "unrelated" condition, the same target (*boy*) was preceded by a prime that was not semantically related (*boulevard*). Primes were presented in uppercase and targets in lowercase. The condition manipulation was conducted within-participants.

Participants first saw a fixation cross that was displayed for 350ms. It was then followed by 400ms of a blank screen. After that, the prime word appeared for 150ms. The presentation of the prime word was followed by another 650ms blank screen. In other words, the stimulus onset asynchrony (SOA) time, i.e., the time between the offset of the prime word and the onset of the target word was 650ms. Finally, participants saw the target word, which they had to make the lexical decision on. If a participant did not make a lexical decision within 3000ms of the onset of the target, the experiment moved on to the next trial.

As mentioned, Experiment 1 was a replication experiment, where I adapted the materials from Thomas et al. (2012). Specifically, the related condition in Experiment 1 used 60 prime-target pairs from Thomas et al.'s "symmetrical associates" (see Table A1 in their Appendix A). These pairs are called symmetrical because the prime has a meaning that evokes the target, and this is also true vice versa, e.g., *girl-boy*, *circle-square*, *salt-pepper*, etc. In the unrelated condition, the same target items were preceded by words that did not have a similar meaning to the target or the related prime. These words were randomly selected from the "forward associates" in Thomas et al.'s Table A1 (Appendix A).

In addition to the 60 experimental items, the experiment also included 60 fillers items, where targets were non-words. Of the fillers, 30 were 4-10/11 letter pseudohomophones that I generated from the ARC Nonword Database (Rastle et al., 2002) —e.g., *spraized, knewed* —, and 30 were non-words from Lupker and Pexman's (2010) Appendix A Standards-1 —e.g., *cleam, dronk*. The experiment started with 10 practice items: 5 words and 5 non-words. For the first 4 practice items, participants saw reminder labels on the screen that the F key corresponded to non-word and the J key to word. These reminders were not present for the latter 6 practice trials, or for the main experiment.

2.2.2 Hypothesis and predictions

Given that Experiment 1 is a replication study, I predict to find the same effects as Thomas et al. (2012). Namely, I predict shorter reaction times (RT) in the lexical decision task in the related, as compared to the unrelated condition. The target that is to be recognized as a word of English is the same in both conditions, but importantly, in the related category, it has been preceded by a word that is similar in meaning. This related prime should activate the meaning of the target, and therefore facilitate its recognition. In the unrelated condition, the prime word would not activate the target, and therefore the target should be recognized at a "baseline" speed, related to its frequency, length, and other properties. (See also i.a., Swinney 1979; Swinney et al. 1979 for classic findings of semantic priming.)

2.2.3 Results and discussion

Data points with incorrectly answered lexical decision responses (i.e., a "non-word" response) were excluded, which removed 2.09% of the data. Figure 2.1 shows the results of Experiment 1: mean RT by condition. For the statistical analysis, a linear mixed effects regression model was fit, using the lme4 package in R (Bates et al., 2015). The model predicted RT on the target word by Condition ("related" vs. "unrelated"). The fixed effects predictor Condition was sum-coded (related: -0.5 and unrelated: 0.5). Random intercepts and slopes were included for participants and items. RTs in the related condition were found to be significantly faster than in the unrelated condition (Estimate=25.51, Std. Error=8.65, t=2.95, p<0.01). That is, participants recognized words faster when they have been primed by a semantically related word.

This finding successfully replicates Thomas et al. (2012). Importantly, they carried out priming experiments in the lab, while the current Experiment 1 relied on web-based data collection. It must be noted, however, that even though I replicated the semantic priming effect, the magnitude of this effect is smaller than what was found by Thomas et al. (2012). In their experiment, the mean RT in the related condition was 531ms, and in the unrelated condition it was 583ms. (See their Table 2 on p. 629). This mean that the numerical magnitude of the facilitation effect was larger in their

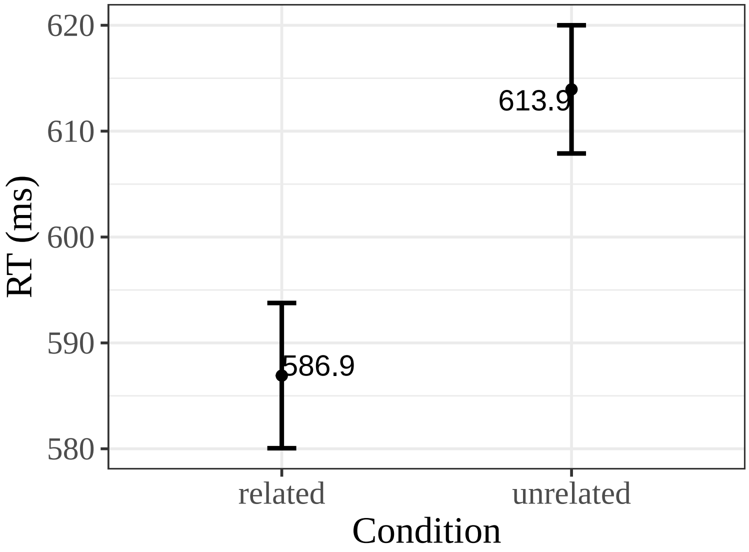


Figure 2.1: Results of Experiment 1: replication of Thomas et al. (2012)

experiment (51ms) than in my Experiment 1 (27ms).

Altogether, though, successfully replicating a previous semantic priming effect on the web validates the subsequent priming experiments that investigate scalar alternatives (Experiments 2-4).

2.3 Experiment 2: lexical priming

Experiment 2 was another (lexical) semantic priming experiment. It tested whether weaker scalar terms from a scale (e.g., *some, good*) prime stronger alternatives (e.g., *all, excellent*). Importantly, scalar terms were not placed in a sentential context, where scalar inference could have been calculated. This means that if we see semantic priming between scalar terms in Experiment 2, that will be because of their semantic similarity, and not because of SI calculation. Experiment 2 therefore provides a baseline for later experiments that test inference-triggering sentences.

2.3.1 Participants and task

49 native speakers of American English participated in an online (PCIbex) experiment. Participants were compensated $2. Participant recruitment and screening was identical to Experiment 1, including the exclusion criterion. Data from 44 participants is reported below.

Capitalizing on the scalar diversity phenomenon (i.a. van Tiel et al. 2016), our testing ground for the activation of alternatives is 60 lexical scales (adjectival, verbal, adverbial and quantifier). The corpus work carried out to collect the 60 lexical scales will be described in more detail in Section 4.2 (Chapter 4). Participants were first presented with a weaker scalar term such as *good*. They then saw the scalar alternative *excellent*, and had to indicate by button press whether this word was a word of English or not. This experimental condition is referred to as "related". In the "unrelated" condition, participants were first presented with a word that was unrelated to the lexical scale, e.g., they saw *foreign* before making a lexical decision on *excellent*. Unrelated words were generated to satisfy two criteria. First, they had to fit into sentence frames that were employed in Experiments 3 and 4, e.g., given the sentence *The movie is good*, *foreign* was chosen, since *The movie is foreign* is also an acceptable sentence. Second, unrelated primes had to have sufficiently low semantic similarity with the target. Semantic similarity was operationalized using vector semantics, which will be described in detail in Section 4.4.5, and in particular using the GLoVe model. (I also discuss semantic similarity in more detail in Section 2.6, when I compare semantic similarity between prime-target pairs in Experiments 1 and 2.)

Other than the critical test items, Experiment 2 was identical to Experiment 1 in its task, procedure (including timing parameters such as SOA), filler and practice items.

2.3.2 Hypothesis and predictions

If pairs of scalar terms (e.g., *good-excellent*) are similar enough in meaning to result in semantic priming, then the results of Experiment 2 should pattern similarly to Experiment 1. Than is, we should see facilitated reaction times in the related condition, as compared to the unrelated condition.

2.3.3 Results and discussion

Data points with incorrectly answered lexical decision responses (i.e., a “non-word” response) were excluded, which removed 2.35% of the data. Figure 2.2 shows the results of Experiment 2: mean RT by condition. Statistical analysis was identical to Experiment 1, with the only difference being in the random effects structure: random intercepts were included for participants and items, and random slopes were included for participants. The statistical analysis revealed no significant difference between RTs in the related and unrelated conditions (Estimate=11.46, Std. Error=9.94, *t*=1.15, *p*=0.26).

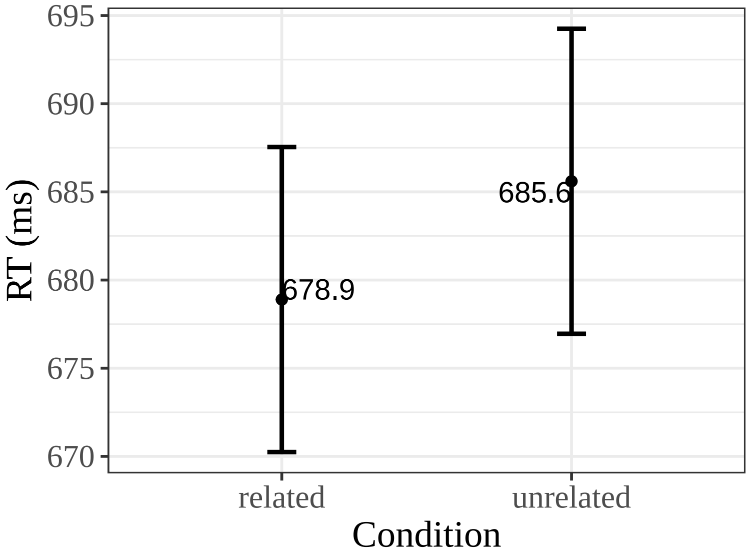


Figure 2.2: Results of Experiment 2: lexical priming experiment testing scalar alternatives

That is, targets in the related condition were not recognized significantly faster than in the unrelated condition. This suggests that pairs of scalar terms do not lead to semantic priming when the words are presented in isolation, in the absence of any sentential context. Therefore, we will be able to conclude that any priming effect we find in sentential experiments (Experiments 3-4) is due to inference processing and alternative retrieval, not just mere meaning similarity.

2.4 Experiment 3: sentential priming

Having seen in Experiment 2 that weaker scalar terms do not prime their stronger alternatives due to meaning similarity, Experiment 3 turns to priming effects in sentential contexts. Specifically, I test whether stronger scalar alternatives are retrieved and activated in the processing of sentences that lead to SI calculation, and are taken to involve reasoning about alternatives.

2.4.1 Participants and task

50 native speakers of American English participated in an online (PCIbex) experiment. Participants were compensated $3.20/3.50. Participant recruitment and screening was identical to Experiment 1, including the exclusion criterion. Data from 46 participants is reported below.

Similarly to Experiments 1-2, Experiment 3 was also a lexical decision task with two within-participants conditions (related vs. unrelated). Target words were the same scalar terms as in Experiment 2. Importantly, however, primes were now full sentences: the prime words from Experiment 2 were embedded in a sentential context. That is, while for the <*good*, *excellent*> scale Experiment 2 used the word *good* as a prime, in Experiment 3 *good* appeared in a sentence: *The movie is good*. As before, the target was *excellent*. Similarly, in the unrelated condition, the unrelated words were embedded in a sentential context, i.e., participants saw *The movie is foreign* before making a lexical decision on *excellent*. Prime sentences were presented word-by-word.

Each trial started with a fixation cross that was displayed for 350ms. It was then followed by 400ms of a blank screen. After that, sentences were presented word-by-word, with each word being displayed for 350ms. There was a 650ms SOA between the offset of the final word in the sentence (*good*/*foreign*), and the onset of the target word (*excellent*). Similarly to previous experiments, if a lexical decision was not made within 3000ms of the onset of the target, the experiment moved on to the next trial.

Filler and practice targets used the materials of Experiments 1-2, but the primes were sentences, not single words.

2.4.2 Hypothesis and predictions

If lexical scalar alternatives like *all* and *excellent* are reasoned about, and retrieved in the process of SI-calculation, then we should see facilitated reaction times in the related condition, as compared to the unrelated condition. That is, *excellent* should be recognized faster when it follows an SI-triggering sentence where it serves as a stronger alternative, than when it follows an unrelated sentence. On the contrary, if lexical alternatives do not play a role in the processing of SI, then there should be no difference in reaction times between the related and unrelated conditions.

2.4.3 Results and discussion

Data points with incorrectly answered lexical decision responses (i.e., a "non-word" response) were excluded, which removed 1.7% of the data. Figure 2.3 shows the results of Experiment 3: mean RT by condition. Statistical analysis was identical to Experiment 2. RTs in the related condition were found to be significantly faster than in the unrelated condition (Estimate=21.62, Std. Error=8.65, t=2.5, p<0.05): targets were recognized faster following an SI-triggering sentence.

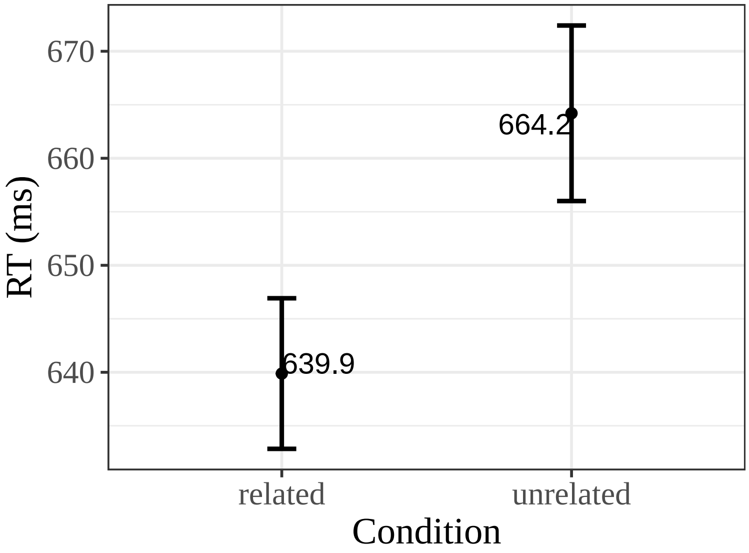


Figure 2.3: Results of Experiment 3: sentential priming experiment testing scalar alternatives

Experiment 3's findings therefore show that a stronger scalar alternative like *excellent* is recognized faster as a word of English when it has been preceded by a sentence that can trigger the *not excellent* SI, namely *The movie is good.* This, in turn, suggests that in the processing of such an SI-triggering sentence, comprehenders retrieved and activated the relevant stronger scalar alternative. In the unrelated condition, on the other hand, such alternative targets would not have been activated in the processing of the prime sentence, and were therefore recognized with a baseline RT. Let us recall that these findings cannot receive an explanation simply in terms of semantic similarity. The prime sentences were identical across the related and unrelated condition up until the critical word (*The movie is X*). And as for the critical word (the weaker scalar term *good* vs. the unrelated word *foreign*), Experiment 2 demonstrated that their difference in meaning, and the similarity between *good* and *excellent* does not, in itself, lead to semantic priming.

There is, however, one important caveat to the above conclusion: we cannot be certain that the priming effect in the related condition shows that the specific lexical item (e.g., *excellent*) has been retrieved. It is also possible that the observed facilitation in RTs is due to a more general activation of semantic features associated with the stronger alternative state. For instance, upon reading the sentence *The movie is good*, perhaps participants considered the stronger alternative state where the movie is more than good, but without necessarily reasoning about the specific alternative *excellent* —this might still result in the observed facilitation in RTs. We also cannot be certain that participants in Experiment 3 actually calculated the SIs (e.g., *not excellent*), since the experiment did not include a task to probe SI calculation. Additionally, as will be discussed in Section 2.6.2, the magnitude of priming across different items does not correlate with the likelihood of SI calculation found in a separate experiment (Experiment 7 in Chapter 4). These issues warrant caution in drawing too strong a conclusion about whether the Experiment 3 findings necessarily suggest that specific lexical scalar alternatives were activated.

2.5 Experiment 4: sentential priming with *only*

As another baseline to Experiment 3, I conducted a sentential priming experiment where prime sentences also included the focus particle *only*. As reviewed in Section 2.1.1, there has been more work on focus alternatives, providing evidence that such alternatives are activated in sentence processing. It will therefore be informative to compare Experiments 4 and 3.

2.5.1 Participants and task

50 native speakers of American English participated in an online (PCIbex) experiment. Participants were compensated $3.20. Participant recruitment and screening was identical to Experiment 1, including the exclusion criterion. Data from 43 participants is reported below.

Experiment 4 was identical to Experiment 3 in its task and procedure (including timing parameters). The critical difference was that in Experiment 4, experimental trials were modified such that prime sentences in the related condition also included the focus particle *only*. That is, before participants made a lexical decision on a stronger alternative target such as *excellent*, they saw the prime sentence *The movie is only good* (presented word-by-word). The unrelated conditions, as well as filler and practice items were entirely identical to Experiment 3, i.e., they were not modified to include the word *only*.

2.5.2 Hypothesis and predictions

The exclusion of alternatives in sentential focus is encoded in the semantics (Rooth, 1985, 1992), while alternatives in SI are excluded pragmatically, in a cancellable way. Given that Experiment 3 already revealed alternative activation for alternatives in SI, we can make a strong prediction that we should see similar effects in Experiment 4. Moreover, this is also what is predicted based on previous work that has tested focus alternatives in a variety of experimental paradigms (see Section 2.1.1). All in all, I make the strong prediction that we should find facilitated reaction times to targets in the related condition, as compared to the unrelated condition.

2.5.3 Results and discussion

Data points with incorrectly answered lexical decision responses (i.e., a “non-word” response) were excluded, which removed 1.98% of the data. Statistical analysis was identical to Experiment

2. Figure 2.4 shows the results of Experiment 4: mean RT by condition. RTs in the related condition were found to be significantly faster than in the unrelated condition (Estimate=24.47, Std. Error=8.01, *t*=3.06, *p*<0.01).

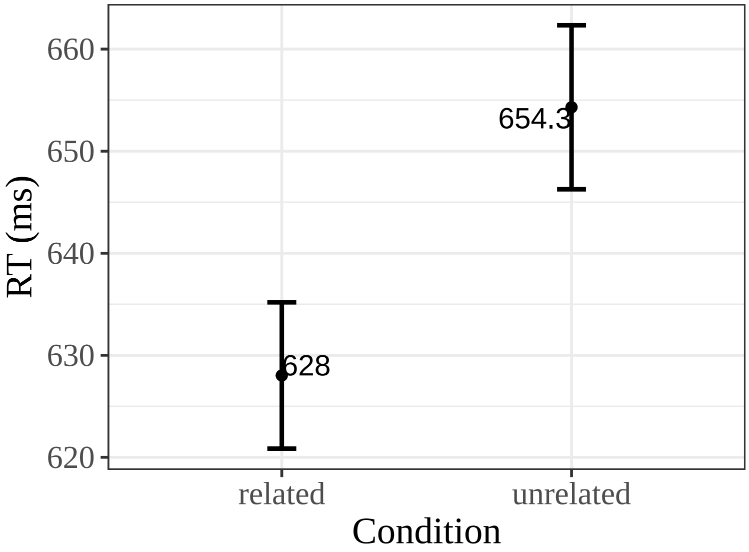


Figure 2.4: Results of Experiment 4: sentential priming experiment testing focus alternatives

Similarly to Experiment 3, Experiment 4 also revealed facilitation for stronger alternative targets in the related condition. That is, the prime sentence *The movie is only good* led to a faster recognition of the word *excellent*, as compared to the prime sentence *The movie is foreign*. Semantic theory holds that sentences including focus (signalled e.g., by *only*) encode the exclusion of alternatives such as *excellent*. (Though there can be other alternatives instead of, or in addition to, the ones that are stronger scalar alternatives like *excellent* —a point that I return to in Chapter 5, Section 5.3.3.) Experiment 4's findings suggest that comprehenders indeed activated such alternatives —in line with existing work on the processing of focus alternatives reviewed in Section 2.1.1.

2.6 Summary of findings

Figure 2.5 shows the results of Experiments 1-4. To reiterate, in Experiment 1 I successfully replicated a known semantic priming effect, thereby validating the web-based priming setup. In

Experiments 3 and 4 respectively, we saw that stronger alternative targets are recognized faster as words of English when they are primed by a sentence whose meaning excludes them as an alternative, either via SI or sentential focus. And in Experiment 2, I ruled out the possibility that the observed facilitation effects between words from the same scale are merely due to meaning similarity.

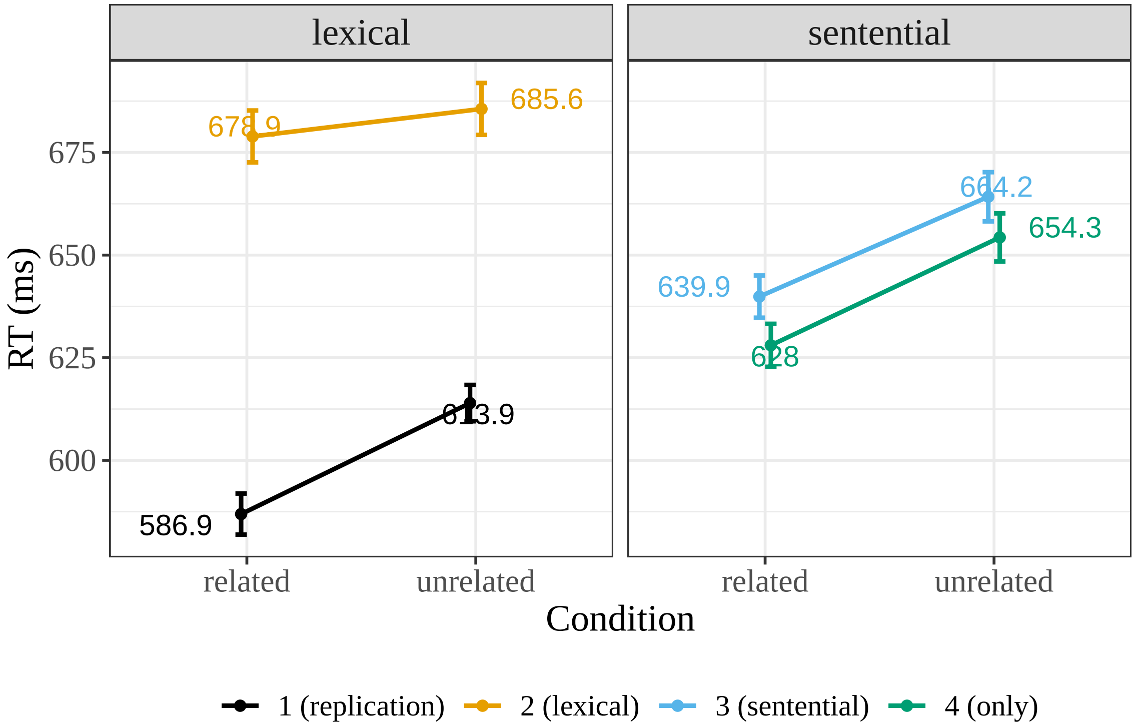


Figure 2.5: Results of Experiment 1-4

The experiments reported in this chapter, and in particular Experiment 3, show evidence that lexical alternatives (*all*, *excellent*) are retrieved and activated in the real-time processing of SI-triggering sentences. This informs our understanding of the mental representations behind pragmatic reasoning. Classic Gricean accounts of SI hold that comprehenders reason about, and derive the negation of, relevant informationally stronger alternatives that the speaker could have said, but did not say. My experimental findings suggest that this abstract reasoning also has processing correlates: relevant alternatives are activated when hearers process SI-triggering sentences.

At the same time, as mentioned in the discussion of the results of Experiment 3, it is still an open possibility, given the current findings, that participants did not retrieve particular lexical alternatives (*all*, *excellent*) per se. Perhaps the priming effects in Experiments 3-4 were not due to the activation of the full lexical representations, but to the activation of some of the relevant semantic features. Future research should further investigate this question, for instance by combining priming tasks with tasks that measure SI calculation, or by using phonological priming (see Section 2.6.3 below).

2.6.1 Relevance for theory?

In this section, I briefly discuss whether the finding that scalar alternatives are retrieved in the processing of SI, or in other words, that lexical scales are psychologically real, can be informative regarding different theoretical accounts of SI. Recall that Neo-Gricean accounts assume that SI calculation proceeds via reasoning about alternatives that are lexically or grammatically specified. A Post-Gricean account such as Relevance Theory, on the other hand, takes SI calculation to be a contextually driven process, and does not attach special importance to lexical scales. This chapter has provided evidence for the retrieval and activation of scalar alternatives in SI processing. One seemingly straightforward way to interpret these findings is that they support Neo-Gricean accounts of SI, since those take hearers to reason about particular lexical alternatives. We could also say that such results are not predicted by theoretical accounts of SI that dispense with lexical scales, such as Post-Gricean accounts. However, as I will argue below, at least two issues arise with this interpretation of the results, having to do with what predictions different theories of SI may make for priming data.

First, it is not clear whether Neo-Gricean accounts would predict the activation of stronger lexical alternatives upon the presentation of weaker scalar terms for that scale, when those weaker scalar terms do not occur in the context of a sentence. One the one hand, we could assume that lexical scales should only be relevant in language processing when alternatives are actually reasoned about, in the context of an SI-triggering utterance. If this is the case, then the findings of this chapter do indeed support Neo-Gricean accounts, since we found priming in a sentential context

(Experiment 3), but not when words were presented in isolation (Experiment 2). On the other hand, if lexical scales are hardwired into, and have a special status in the lexicon, then perhaps we should predict that pairs of scalar terms prime each other even in the absence of a sentence that would lead to SI. As mentioned, de Carvalho et al. (2016) made a prediction along these lines for Neo-Gricean accounts (though more specifically claiming that the weaker member of a lexical scale primes the stronger member asymmetrically). If we are to follow this reasoning, then the findings of Experiments 2-3 in fact do not fully support Neo-Gricean accounts, since no priming effect was found when the scalar terms occurred on their own (Experiment 2). The finding that the activation of alternatives only showed up in a (sentential) context (Experiment 3) could even be argued to better support Relevance Theory, which takes SI calculation to only occur when there are sufficient supportive cues from context.

Second, another reason why it is not trivial to link priming evidence to theoretical accounts of SI is that the activation of alternatives may signal either that alternatives were retrieved for the SI calculation process to occur, or we might see activation as a by-product of the SI calculation process. In other words: is there SI calculation because of priming (i.e., alternative activation), or is there priming because of SI calculation? I elaborate in the following. Broadly speaking, Neo-Gricean accounts would assume that for SI to arise, particular lexical items from lexical scales are reasoned about —this would predict the priming effect we found in Experiment 3. But at the same time, it is also possible that the priming effect we see is epiphenomenal. On Post-Gricean theories of SI, hearers still calculate SI, even though lexical scales do not play a special role in this process. And once hearers have reached the SI-enriched meaning (≈*The movie is no more than good.*), perhaps this can then lead to the observed priming effect, even if the stronger alternative *excellent* was not retrieved in the first place. These two possibilities of interpreting the priming findings are related to the issue I have discussed above: whether facilitated RTs constitute evidence that a specific lexical item *excellent* was retrieved, or whether they simply suggest that some semantic features related to the alternative state “the movie is more than good” were retrieved.

Given all this, I would argue that as things stand, no firm conclusions can be reached about the validity of Neo-vs. Post-Gricean accounts based on priming evidence. More research is needed to precisely pin down what processing predictions can be made from different theoretical accounts, and which account the existing empirical data therefore supports.

2.6.2 Remaining empirical puzzles

In this section I discuss a number of empirical puzzles and open questions that emerge from Experiments 1-4.

First, we saw that Experiment 3 and 4 pattern alike: in both experiments, RTs were facilitated when the stronger scalar alternative target was preceded by its weaker scalemate. Additionally, I conducted a statistical analysis on the combined Experiment 3-4 data set. A linear mixed effects model was fit, predicting RT by Condition (related vs. unrelated), Experiment (Experiment 3 vs. 4) and their interaction. The fixed effects predictors were sum-coded: within Condition, related: -0.5 and unrelated: 0.5, and within Experiment, Experiment 4: -0.5, and Experiment 3: 0.5. Random intercepts were included for participants and items, as well as random slopes for participants for the Condition predictor. This statistical model revealed a significant effect of Condition (Estimate=22.33, Std. Error=5.98, t=3.73, p<0.001), but not of Experiment (Estimate=9.51, Std. Error=22.53, t=0.42, p=0.67) or the interaction (Estimate=-5.45, Std. Error=11.91, t=-0.46, p=0.65). In other words, both experiments (analyzed on their own or combined) reveal that targets are recognized faster in the related condition than in the unrelated condition, and there is no significant difference between the two experiments.

This suggests that alternatives like *excellent* are similarly activated no matter whether the sentence that is processed is *The movie is good* or *The movie is only good*. This is despite the fact that in the case of SI, the exclusion of alternatives is in the pragmatics, but in the case of focus, it is in the semantics, and one might predict that the latter serves as a stronger cue. Additionally, as will be demonstrated later in the dissertation (Experiment 7 in Chapter 4 and Experiment 13 in Chapter 5), the rate of inference calculation for SI-triggering sentences and sentences with *only* is significantly different. Specifically, an utterance of *The movie is only good* is more likely to lead to the inference *The movie is not excellent* than an utterance of *The movie is good.* The lack of a difference between Experiment 3 and 4 suggests that the activation of alternatives, as measured via priming, does not track the rate of inference from the corresponding sentences: it is not the case that more robust inference calculation corresponds to stronger priming.

Second, Experiment 2 was conducted to rule out the possibility that a priming effect seen in Experiment 3 would reflect mere meaning similarity, rather than the processing correlate of reasoning about alternatives. For this reason, it is a welcome result that Experiment 2 revealed no significant effect. But it is itself a puzzle why this was the case, i.e., why we found semantic priming in Experiment 1 but not in 2. One possibility to explore is whether Thomas et al.'s (2012) items are more similar in meaning than the pairs of scalar terms tested in my Experiment 2. For this, I calculated vector semantic similarity between prime and target in the related condition vs. the unrelated condition. This was done using the GLoVe model and specifically the spaCy word embeddings —a method that will be discussed in more detail in Section 4.4.5. Table 2.1 shows the results, averaged over items: in the related condition, we have a similarity measure between words such as *girl-boy* (Experiment 1) and *good-excellent* (Experiment 2), while in the unrelated condition we have the similarity of word pairs like *boulevard-boy* (Experiment 1) and *foreign-excellent* (Experiment 2).

Cosine similarity	**Related condition**	**Unrelated condition**
Experiment 1 (replication)	0.605	0.126
Experiment 2 (lexical)	0.707	0.138

Table 2.1: Average semantic similarity between prime-target pairs in the related and unrelated conditions in Experiments 1 and 2.

Based on Table 2.1, there does not seem to be a substantial difference between Experiments 1 and 2 in terms of the relevant semantic similarity differences. It seems, then, that the null result in Experiment 2 is not due to the scalar items being less similar to one another than Thomas et al.'s items, at least in the vector semantic sense. It is perhaps possible that the nature of the relationship between prime and target is different in Thomas et al. (2012) (and classic studies on semantic priming) than in the case of scalar items: there is an intuitive sense in which *girl-*

boy, *minimum-maximum* and *salt-pepper* are closer semantically than *good-excellent* and *some-all*. Future research should address what exactly drives semantic priming, and why we did not find it in Experiment 2.

The lack of a priming effect in Experiment 2 is, in a sense, a failure to replicate de Carvalho et al. (2016). But importantly, there are many potentially critical differences between their experiment and mine: e.g., number and identity of the lexical scales tested, the presentation time of the prime (de Carvalho et al. used subliminal priming), as well as conducting the experiment in the lab vs. on the web. Given such differences between the items and methods, it is perhaps not too concerning that the two experiments did not yield converging results, but in the future it will be important to pin down why exactly the difference in results arises.

Third, there is no by-item correlation between the priming effect in Experiment 3 and the rate of SI calculation for the corresponding prime sentence. As we will see in Experiment 7 (Chapter 4), lexical scales vary in how robustly they lead to SI calculation: e.g., the *allowed* → *not obligatory* SI is more likely than *dirty* → *not filthy* (i.a., van Tiel et al. 2016). This, however, does not correspond to a systematic difference in priming. One possible reason could be that in Experiment 3, the priming effect is measured by comparing to the unrelated condition (RT on *excellent* given *The movie is good* vs. *The movie is foreign*.) The specific word in the unrelated condition (here, *foreign*) might introduce variation across items, and given that it is this "varied" range of unrelated conditions that we compare our related condition to, a potential by-item effect could be obscured. One solution in future work could be to use a more uniform control (unrelated condition), relying perhaps on identity priming (*The movie is good*) or antonym priming (*The movie is bad*). Lastly, it is also possible that Experiment 3 lacks sufficient statistical power to reveal a by-item correlation effect. Participant numbers for Experiments 1-4 were determined based on a power analysis conducted for Experiment 1 (using the simr package in R; Green and MacLeod 2016). This suggested that with 40 participants, the experiment is already at/above 80% power. But importantly, the power analysis was only conducted to look at the related vs. unrelated condition effect, not any by-item effects, so it is possible that more data collection would be necessary for those effects to show up.

2.6.3 Future work

In addition to the puzzles discussed in the previous section, there are some other promising avenues of future work, following up on the findings of Experiments 2-4. First, it would be interesting to combine the semantic priming experiment with phonological priming: e.g., to see if a sentence such as *The movie is good* activates a word like *exercise*, which is phonologically similar to the stronger alternative *excellent*. If such activation turned out to be present, that would provide strong evidence that the specific lexical item *excellent* is retrieved in the processing of SI, as opposed to, e.g., a more general concept of “more than good”. Second, it is important to remember that the experiments reported in this chapter all employed a 650ms SOA time, which counts as somewhat long. Previous work on the activation of focus alternatives has found that differences in SOA can be very important (i.a., Husband and Ferreira 2015; Gotzner 2017; Gotzner and Spalek 2019 and references therein). The emerging finding seems to be that early in the time course of processing, with a short SOA, activation can be found for all semantically similar alternatives. It is later in processing, detectable with longer SOAs, that only the relevant alternatives remain activated. My findings seem in line with work on focus alternatives: relevant alternatives are activated even after a long SOA. But in future work, it will be important to directly manipulate the SOA in experiments on scalar alternatives.

CHAPTER 3

PROCESSING COST: CONTEXT OR ALTERNATIVE RETRIEVAL?

A central question that has been asked about pragmatic inferences such as scalar inference concerns their processing: does generating them incur processing cost? An additional question that arises is: if SI calculation is costly, what aspect of the inferential process causes the processing cost? In this chapter, I test two different explanations of processing cost. As we saw in Chapter 1, two crucial properties of SI are that it makes reference to alternatives, and that it is context-sensitive. In prior work, the prediction has been made that reasoning about lexical alternatives is what causes cost (i.a., van Tiel and Schaeken 2017), and it has also been argued that processing cost should track the context (Degen and Tanenhaus, 2014). Here, I test these (potentially competing) hypotheses on two pragmatic inferences.

In addition to SI, the empirical case study in this chapter is *it*-cleft exhaustivity (1):

(1) It is a cookie that Mary ate.
Literal: Mary ate at least a cookie.
Inference-enriched: Mary **only** ate a cookie.

In the case of (1), just like in SI, hearers regularly go beyond the literal meaning of the sentence and calculate the inference-enriched meaning. This suggests that, at least on a descriptive level, *it*-cleft exhaustivity is similar to scalar inference. The motivation for including *it*-cleft exhaustivity in my empirical scope is to probe whether (and to what extent) the context-sensitivity of pragmatic inferences extends beyond SI. Additionally, as we will see, it has been argued that the exhaustivity inference in *it*-clefts is not the result of alternative-based reasoning, contrasting with SI, and making these two inferences a useful case study for testing whether alternative retrieval is what leads to processing cost.

3.1 Background

In this section I first review different theories regarding whether the processing of implicatures is costly (Section 3.1.1). I then discuss existing theoretical and experimental work on the context-sensitivity of SI calculation (Section 3.1.2). This leads to two different proposals regarding the source of the processing cost of implicature calculation: context vs. Lexical Access. I review these proposals in Section 3.1.3, and the experiments reported in this chapter are aimed at teasing them apart, using SI and *it*-cleft exhaustivity as testing grounds.

3.1.1 The cost of implicature processing

One of the major questions in psycholinguistic studies of semantics-pragmatics is how fast implicatures are processed: are they processed on par with other inferences, i.e. does generating them incur a cost? This is meant to inform our understanding of the relations between semantic and pragmatic processes during language comprehension. In the following I review three different approaches to this question.

Default hypothesis

Levinson (2000)'s default hypothesis takes implicatures to be default inferences, derived automatically and regardless of context. This predicts implicature calculation to be immediate and effortless, where inference-enriched interpretations will always precede and be accessed faster than literal interpretations. On the contrary, it is cancelling a conversational implicature that requires processing resources, and therefore time. Experimental evidence for the default hypothesis comes i.a. from Grodner et al. (2010), who found eye-tracking evidence for a rapid interpretation of the inference-enriched meaning of *some*. Nonetheless, though Grodner et al. (2010) showed that scalar inferences are derived without processing delay, a further prediction of the default hypothesis, namely that the derivation of the literal interpretation (*at least some*) should be associated with a processing delay, has not been confirmed.

Literal-first hypothesis

Others assume a two-stage processing sequence: literal interpretations are necessarily computed before, and therefore accessed faster than inference-enriched ones (i.a. Huang and Snedeker, 2009). This hypothesis is often referred to as the literal-first hypothesis (Degen and Tanenhaus, 2014; Breheny, 2019). Though scalar implicature processing may be rapid, it has to be preceded by some semantic analysis – in line with a model of linguistic architecture that takes semantic representations to serve as a mediator between phonology and pragmatics. A large body of research has indeed shown that implicature generation is costly, as evidenced by increased reaction times (Bott and Noveck, 2004), ERP patterns (Noveck and Posada, 2003), or delays in eye-tracking (Huang and Snedeker, 2009).

Probabilistic frameworks

Under the default and literal-first hypotheses, semantics and pragmatics are clearly separated in the grammar and processing. Probabilistic frameworks, on the other hand, do not posit such a sharp boundary; rather, semantic and pragmatic processes are taken to be intertwined. Degen and Tanenhaus (2014)'s constraint-based framework (see also Degen and Tanenhaus, 2019 for an overview), for example, does not assign a privileged status to either literal or inference-enriched readings – neither reading is taken to always require more/less processing resources or time than the other. Instead, it is assumed that how fast a scalar implicature is computed is in large part determined by context. Robustness of inference calculation, as well as speed and ease of processing, depend on the strength of cues available. The more probabilistic support there is from such cues, the faster comprehenders will arrive at the inference, and the less easily cancellable that inference will be.

The primary goal of these probabilistic models is not to test whether implicatures are calculated by default or at a cost, but rather to identify and quantify the cues that hearers rely on when generating inference-enriched meanings. Cues that have been shown to have an effect on the rate or cost of implicature calculation are e.g. the syntactic partitive, or the availability of lexical alternatives (Degen and Tanenhaus, 2014); the relevance of the stronger alternative proposition

(Breheny et al., 2006; Politzer-Ahles and Fiorentino, 2013); cognitive load (De Neys and Schaeken, 2007); the speaker's knowledge state (Goodman and Stuhlmüller, 2013); or face threat (Bonnefon et al., 2009). QUDs are also taken to be such a probabilistic cue, and have thus been predicted to influence implicature calculation and processing (Degen, 2013; Degen and Tanenhaus, 2014). In the following, I turn to previous work on the role of context and QUDs.

3.1.2 Context effects on SI calculation

It has long been noted that, depending on the discourse context, otherwise predicted implicatures can fail to arise (see i.a. Van Kuppevelt, 1996). Following Roberts (1996/2012), I formalize context using the notion of Questions Under Discussion (QUDs), defined as the immediate topic of discussion, which proffer a set of relevant alternatives. Discourse is construed as giving rise to a stack of QUDs, and the ultimate discourse purpose is to answer all of these QUDs. An assertion is felicitous, then, if it chooses among the proffered alternatives and thereby bears upon the QUD. To see how QUDs could modulate implicature calculation, let's consider the example in (2):

(2) a. Mary ate some of the deep dish.
b. Mary ate all of the deep dish.
c. Mary ate some but not all of the deep dish.

Assuming the QUD *How much deep dish did Mary eat?*, *Mary ate some of the deep dish* and *Mary ate all of the deep dish* are both in the set of proffered alternatives. That is, both (2-a) and (2-b) would be felicitous responses to the QUD. The hearer of (2-a) is therefore predicted to identify (2-b) as a felicitous alternative, leading her to realize that the speaker's choice to utter (2-a) implicates the negation of (2-b), ultimately deriving the inference in (2-c).

As the below example from Levinson (2000) shows, QUDs can also discourage the calculation of an implicature:

(3) A: Is there any evidence against them?

B: Some of their identity documents are forgeries.

Dispreferred implicature: Not all of their identity documents are forgeries.

The potentially available but dispreferred implicature in (3) would be consistent with the common ground and B's utterance; however, it is less likely to arise than the implicature in (2), because A's question suggests that she is only interested in whether there is at least some evidence against the criminals. Thus there is no particular reason to consider *All of their identity documents are forgeries* as an alternative that B could have said. We can thus see that QUDs offer a way to capture whether an implicature is more or less likely to arise.

As mentioned, under a probabilistic constraint-based model, QUD is one of many cues that is predicted to influence how likely an implicature is to be calculated. Predictions regarding the effect of QUDs on implicature calculation are supported by ample empirical data in experiments with both adults and children. It has long been observed that children are not adult-like in how they interpret structures that give rise to two or more competing readings, and are therefore more reluctant to calculate implicatures than adults (see e.g. Chierchia et al., 2001; Noveck, 2001). However, Papafragou and Tantalou (2004) showed that in contexts approximating naturalistic conversations, children can and do calculate implicatures. Investigating well-known lexical (*some*-*all*) as well as more context-dependent and ad hoc implicatures, the authors found that children exhibit robust rates of implicature calculation. Similar effects have been observed in the domain of scope ambiguities, where children are known to resort to surface (as opposed to inverse) scope interpretation more than adults (Musolino, 1998, 2011), but actually show adult-like behavior given the right context (Gualmini et al., 2008).

QUDs have also been shown to induce variation in calculation rates in adult interpretations. Experiments have been carried out either using an explicit QUD (Zondervan et al., 2008; Yang et al., 2018), or promoting one implicitly via a background story (Degen, 2013) or via focus intonation (Cummins and Rohde, 2015). Zondervan et al. (2008), Degen (2013) and Yang et al. (2018) investigated *some but not all* scalar inferences, placing them under QUDs containing *some* vs. *all*, *none* vs. *all* and *any* vs. *all*, respectively. They found that reliably more inferences were calculated when an *all*-QUD is promoted. For example, Zondervan et al. (2008) presented participants with the sentence *Some pizzas were delivered*, which had to be evaluated with respect to either a question containing *some* (4), or one containing *all* (5).

(4) A: Were some pizzas delivered?
B: Some pizzas were delivered.

(5) A: Were all pizzas delivered?
B: Some pizzas were delivered.

The study found a 7% calculation rate of the *some but not all* inference in (4), but a 43% calculation rate in (5). Cummins and Rohde (2015) found comparable results testing a wider variety of scalar inferences, e.g. *warm-not hot*, which is especially important in light of recent findings that there is substantial variability among scales such as *some-all* vs. *warm-hot* with regard to the availability of the corresponding implicature (i.e. scalar diversity, van Tiel et al., 2016). This constitutes empirical confirmation that hearers do not always calculate scalar inference; rather, they are more or less likely to do so depending on context.

3.1.3 Processing cost: lexical access or context?

In Section 3.1.1 I discussed different theories about whether pragmatic inference calculation results in processing cost. More specifically, there are also different proposals as to what aspect of the inferential process incurs a cost. Van Tiel and Schaeken (2017) (following Chemla and Bott, 2014) take as their starting point Katzir's (2007) structurally based theory of alternative construction and complexity, which I reviewed in Section 2.6.1 in Chapter 1 (see also Fox and Katzir, 2011). Specifically, the authors propose that the aspect of implicature calculation that causes a delay is lexical retrieval during alternative construction (Lexical Access hypothesis[1]). For example, in

[1] Lexical Access is of great contemporary interest not only as a proposed source of processing cost when calculating pragmatic inferences, but also to capture child language data. In particular, recent work has argued that children's

scalar inference, to construct the alternative *Mary ate all of the deep dish* for *Mary ate some of the deep dish*, the lexical item *all* needs to be retrieved, which will incur a processing cost. On the other hand, the calculation of *it*-cleft exhaustivity is argued to proceed without recourse to lexical alternatives, predicting no processing cost. In addition to SI and *it*-cleft exhaustivity, van Tiel and Schaeken (2017) also investigated conditional perfection and free choice inference, exemplified below.

(6) Conditional perfection
Target sentence: Each of the shapes is red if it is a circle.
Inference: Not all of the shapes are red.
Alternative: Each of the shapes is red.

(7) Free choice inference
Target sentence: Each of the shapes is red or green.
Inference: Not all of the shapes are red/green.
Alternative: Each of the shapes is red/green.

For free choice and conditional perfection, the alternatives are a subset of the target sentences and can therefore be constructed via deletion, e.g., the alternative *Each of the shapes is red* can be arrived at by deleting two words from the target sentence *Each of the shapes is red or green*. Of the four pragmatic inferences, then, only SI should incur a processing cost, because only in that case is lexical access needed for alternative construction. Van Tiel and Schaeken (2017) employed a sentence-picture verification task, comparing reaction times to an unambiguous sentence with reaction times to a target sentence on its inference-enriched reading. Their empirical findings were in line with the Lexical Access hypothesis: it is only scalar inference (*some-all*) that resulted in processing effort, the other three inference types did not.

Importantly, there are also other hypotheses about what can lead to processing cost in prag-

divergence from adult behavior when it comes to the interpretation of scalar inference or disjunction is due to their inability to access lexical alternatives (see i.a. Barner et al., 2011; Singh et al., 2016) matic inference calculation. In line with the general context-sensitivity of SI, constraint-based models predict QUDs to influence not only the likelihood of implicature calculation, but also the processing cost associated with this calculation (Degen, 2013; Degen and Tanenhaus, 2014). This predicts no uniform cost or lack of cost for calculating conversational implicatures per se. Instead, it attributes particular importance to supportive vs. non-supportive contexts, and predicts that the likelihood of inference calculation, and crucially also whether this calculation causes a delay in reaction times, is a function of how much the target inference is supported by the QUD. When a target inference is congruent with the QUD, an increased rate of inference generation is predicted, as well as decreased reaction times. When the target inference is not supported by the QUD, the opposite pattern is predicted.

Indeed, Degen (2013) and Degen and Tanenhaus (2014) put forward the prediction that a QUD that makes the alternative *all* more relevant should lead to faster calculation of scalar inference than one that makes *none* salient. Degen (2013) explicitly manipulated the QUDs to test these predictions, and found that calculating the scalar inference is numerically faster under an *all*-QUD than under an *any*-QUD (Experiment 2a). Moreover, Degen (2013) and Degen and Tanenhaus (2014) also observed individual differences for scalar inference calculation, such that some participants consistently calculated the inference, some consistently did not, while a third group was inconsistent. The authors argued that participants' response consistency is indicative of how much uncertainty they had about the QUD. Inconsistent participants were argued to have more uncertainty, leading to a higher cost for generating the inference. Similarly, Kursat and Degen (2020) have found that reaction times are influenced by an interaction between the QUD and 'participant type' (i.e. whether or not a participant tends to calculate the inference), though their results did not reveal an overall modulating effect of QUDs.

3.2 Contributions of Chapter 3

This chapter investigates the processing cost of calculating SI and *it*-cleft exhaustivity, testing the empirical validity of the two proposals reviewed above regarding the source of processing cost. To do so, inference-triggering sentences are placed in different discourse contexts, i.e., under different QUDs. Table 3.1 spells out the predictions of a QUD-based hypothesis (Degen, 2013; Degen and Tanenhaus, 2014), as compared to the particular alternative-based hypothesis tested here, Lexical Access (van Tiel and Schaeken, 2017; Chemla and Bott, 2014). As can be seen, the two proposals make diverging predictions.

		Cost for scalar inference	Cost for *it*-cleft
Lexical Access hypothesis		✓	✓/×
QUD hypothesis	Literal-biasing question	✓	✓
	Inference-biasing question	×	×

Table 3.1: Predictions about the reaction time cost of implicature calculation

If Lexical Access is right, then scalar inference should always incur a processing cost, because the lexical alternative *all* is always retrieved. For *it*-cleft exhaustivity, on the other hand, predictions depend on whether only lexical, scalemate alternatives trigger cost. *It*-cleft exhaustivity can be taken to have ad-hoc contextual alternatives (e.g., *It is a muffin that Mary ate* is an alternative to *It is a cookie that Mary ate*), but not scalemate alternatives as in scalar inference. Thus, if retrieving any kind of alternative is costly, *it*-cleft exhaustivity calculation is predicted to be costly. However, if only the retrieval of scalemate alternatives is costly, then calculating *it*-cleft exhaustivity should not incur a cost, because the lexicon does not need to be accessed in the derivation process. Crucially, regardless of the kind of alternatives one considers for *it*-cleft exhaustivity, an alternative-based account would likely predict uniform (lack of) cost across different contexts.

In contrast to alternative-based hypotheses, the QUD hypothesis predicts that for both scalar inference and *it*-cleft exhaustivity, whether cost is incurred depends on the type of question the target sentence addresses. If a QUD supports the derivation of the inference (Inference-biasing), then there should be no increase in reaction times, but if the QUD biases against deriving the inference (Literal-biasing), processing cost is predicted. Crucially, this means that there should be no general default cost for alternative retrieval: given a supportive context, any alternative-sensitive inference should be calculated without a delay in reaction times.

The empirical findings reported in this chapter are in line with the QUD hypothesis. Specifically, as we will see in Experiment 6, under a QUD that makes the target inference likely to arise, inference computation does not cause a delay in reaction times. On the contrary, under a QUD that makes inference calculation unlikely, that calculation is time-consuming. Additionally, in Experiment 5, I address the known problem that there exists no explicit mechanism for identifying the QUDs relevant for a given context. I take the first step to address this problem, and go beyond previous studies by establishing the relevant QUDs for a given utterance in a more empirically grounded manner, viz. by relying on experimental production data. The assumption that has often been made is that the relevant questions are the ones that contain a member of the given lexical scale, i.e. for the target sentence *Mary ate some of the deep dish*, the possible questions would be *Did Mary eat some/none/all of the deep dish?*. Although the assumption that these are the relevant QUDs is informed by theory, there has not been systematic empirical work probing the possible range of QUDs.

3.3 Experiment 5: QUD elicitation

In this section I present an elicitation experiment, which established the likely QUDs for a given context for two types of pragmatic inferences: scalar inference (SI) and *it*-cleft exhaustivity (EXH).

3.3.1 Participants

40 monolingual speakers of American English participated in an online elicitation experiment, administered on the Ibex platform (Drummond, 2007). Participants were recruited on Amazon Mechanical Turk. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Participants were compensated $2.00.

3.3.2 Task, materials and procedure

The task was a modified sentence-picture verification task used for elicitation. Participants were provided with a background story that two people, Anne and Bob, are discussing pictures about shapes. Anne cannot see these pictures, so she is always asking Bob about what he sees. The instructions also emphasized that Bob always answers truthfully. (8) shows the instructions given to participants before the start of the experiment.

(8) In this experiment you are going to see dialogues between Anne and Bob, who are discussing pictures. Each picture shows a number of colored shapes.
However, only Bob can see the pictures, Anne cannot. So Anne is always asking Bob about what he sees.

Participants then saw written SI and EXH target sentences paired with pictures, and were told that the sentences were Bob's answers to Anne's questions. Target sentences and pictures were adapted from the materials used by van Tiel and Schaeken (2017), with the addition of the context manipulation. The target sentences investigated were of the following form: SI: *Some of the shapes are blue* and EXH: *It is the square that is blue*. In place of Anne's question participants saw a blank line. An example of a trial screen is in (9).

(9) Anne: ____________________________*?*
Bob: *Some of the shapes are blue. / It is the square that is blue.*
Picture: Good Control or Target

The task was question elicitation: participants were instructed to guess what Anne's question could have been, based on Bob's answer and the picture, and had to provide their response in writing. Linguistically overt questions and the QUDs they evoke are different notions (see i.a. Hawkins et al., 2015 and references therein). Nevertheless, this task still gathers a measure of what participants think the target sentence is in response to, and is therefore a good proxy for tracking

QUDs. For ease of explication, I will refer to the stimulus questions throughout as ‘QUDs’, which strictly speaking are the conversational topics tracked by the overt questions.

There were two types of pictures: Bob’s answers were unambiguously good descriptions of Good Control pictures. For Target pictures, they were good descriptions on their literal (SI: *At least some of the shapes are blue*, EXH: *The square is blue*), but not on their inference-enriched reading (SI: *Some but not all of the shapes are blue*, EXH: *Only the square is blue*). Examples of the pictures can be seen in Figure 3.1.

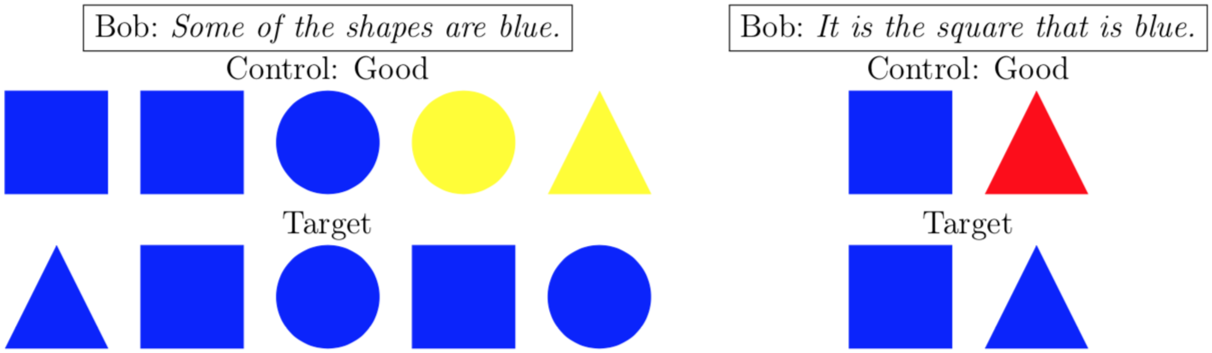

Bob: *Some of the shapes are blue.*
Control: Good
Target
Bob: *It is the square that is blue.*
Control: Good
Target

Figure 3.1: Example experimental trial from Experiment 5: SI (left) and EXH (right)

SI Good Control pictures always contained five shapes, three of which matched the color mentioned in Bob’s sentence (here, blue), while two were a different color (here, yellow). SI Target pictures contained five shapes of the same color, the one mentioned in the sentence. EXH Good Control pictures depicted the shape with the color that was mentioned (here, blue square), as well as a different shape with a different color (here, red triangle). EXH Target pictures depicted two shapes of the mentioned color. Shapes were varied between triangle, circle and square (with SI pictures containing either a mix of shapes or the same shape five times), and colors were varied between blue, yellow, red, green, orange and black.

The experimental design included the two-level factor of Picture type as a between-participants manipulation: 20 participants saw only Good Control, and 20 other participants only Target pictures. A previous pilot with a within-participants design produced qualitatively the same results. Each participant saw 15 SI trials and 15 EXH trials. The experiment was administered in a Latin

Square design.

3.3.3 Results and discussion

Results were coded in the following way. Whenever two responses only differed from each other in the mentioned color or shape, they were coded as the same type of question. For example, *Are any of the triangles yellow?* and *Are any of the shapes red?* were both coded as *any*. Under each question type, I collapsed across closely related linguistic variants, for example *any of the shapes* and *any shapes* were both coded as *any*. In (10)-(11), I demonstrate the most commonly offered QUD types.

(10) Most frequent SI question types
what: *What color are the shapes?*
any: *Are any (of the) shapes blue? Are there (any) blue shapes?*
all: *Are all of the shapes blue?*
some: *Are some of the shapes blue?*

(11) Most frequent EXH question types:
which: *Which/what shape is blue? Which one (of them) is blue?*
any: *Are any of the shapes blue? Are there any blue shapes?*
what: *What color are the shapes? What color is the square?*

Though both SI and EXH elicited thirteen distinct types of questions each, the majority of answers came from a much smaller set for both constructions. Table 3.2 shows the frequencies of the most frequent types of questions in the data. The question types not included here occurred with less than 5% frequency[2].

[2] The question types that each occurred with less than 5% frequency were the following. SI: *What's the dominant color?*, *Which color has the most shapes?*, *What color are some of the shapes?*, *What is one of the colors?*, *How many (shapes) are blue?*, *Which shapes are blue?*, *What is blue?*, *Are a lot of shapes blue?*, *If there are squares, what color are they?*; EXH: *Is one of the shapes blue?*, *What shape is it?*, *Which if any shapes are blue?*, *Is the triangle blue?* (7% frequency with Target), *Is it the circle or the square that is blue?*, *Are both shapes blue?*, *What is the color of the*

	SI				EXH		
	what	any	all	some	which	any	what
Target	42%	25%	6%	12%	54%	9%	8%
Good Control	32%	33%	20%	2%	67%	14%	6%

Table 3.2: Results of Experiment 5: Frequencies of elicited question types

We can see that a small number of questions dominated in the elicitation task for each construction. But for SI, two types seem roughly equally frequent (*any*- and *what*-questions), whereas for EXH one type of question (*which*-questions) is clearly favoured over all others.

Another observation to be made about the data is that for both Good Control and Target pictures, the same type of questions were elicited and in largely the same frequencies. This is somewhat puzzling, considering that several of the elicited questions were previously thought to bias towards one or the other interpretation. For example, *all*-QUDs have been argued to bias towards the enriched reading, and *any*-QUDs towards the literal reading. Based on this argument, we might predict that only Good Control pictures, which are compatible with the inference-enriched meaning of SI, would have elicited e.g. *all*-questions. Conversely, we might not have predicted Target pictures, which are not compatible with the enriched reading, to elicit *all*-questions. One thing to keep in mind, however, is that in this experiment we do not obtain information about what interpretation participants actually assigned to the target sentences. As many studies (including Experiment 6) have found, some participants do judge *Some of the shapes are blue* to be a good description of both Good Control and Target pictures – in these cases, it is perhaps unsurprising that both pictures elicited the same type of questions. Another explanation might be that when participants provided *all*-questions in the Target condition, this was driven by the salience of the picture (where all shapes were blue), and not by question-answer congruence considerations. Future research should further probe these issues about the interplay of production and interpretation.

The primary aim of Experiment 5 was to test in a more systematic and empirically grounded manner what the likely QUDs are for a given dialogue context, where the dialogue includes either a *some* or an *it*-cleft sentence. Previous work on the role of context in SI calculation has largely

shape on the right?, *Which shape is on the left?*, *What shape except for circle is blue?*, *Is it the square that is blue?*.

assumed questions that contain members of the relevant lexical scale, while no investigation to my knowledge has been conducted about EXH in this respect. We can see that the elicitation study has indeed uncovered questions that have not been discussed in existing literature as relevant to implicature derivation or the particular target sentences, e.g. the *what*- and *which*-QUDs. On the other hand, some questions previously discussed in theoretical or experimental contexts did in fact show up in the elicitation data, e.g. *any*-, *some*- and *all*-QUDs.

Experiment 5 established likely QUDs whose effects on SI and EXH computation we can then investigate. I therefore took the most frequent questions for each construction from Experiment 5, and used them in Experiment 6 to see if they modulate implicature calculation rates and processing cost.

3.4 Experiment 6: QUDs modulate calculation rates and processing cost

I present a sentence-picture verification experiment, where I embedded scalar inference and *it*-cleft exhaustivity sentences under the most frequent QUDs elicited in Experiment 5. The results showed that the probability of inference calculation, as well as whether there is an increase in reaction times, is conditioned on the QUD.

3.4.1 Participants

90 (30 in each QUD condition, different from those in Experiment 5) native monolingual speakers of American English participated in the experiment, administered on the Ibex platform (Drummond, 2007). Participants were recruited on Amazon Mechanical Turk. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. Participants with a mistake rate exceeding 25% on filler items were removed from the analysis. This resulted in the removal of one participant from the wh-word experiment and four participants from the quantifier-experiment. Participants were compensated $1.50.

3.4.2 Task, materials and procedure

Experiment 6 employed the sentence-picture verification paradigm, with target sentences embedded in a dialogue context. Participants were given the same background story as in Experiment 5: Anne is asking questions from Bob, about pictures that only Bob can see. On the first screen, participants saw Anne's question, and they were instructed to press a key after they have read it. On the following screen, they saw Bob's answer to Anne's question, as well as the picture they were discussing. Participants were instructed to make a binary judgment (by clicking on a button) about whether Bob gave a good answer to Anne's question, given the picture he saw. The two buttons said 'Good' and 'Not Good'. Participants' choices were recorded, as well as reaction times from the onset of the second screen (displaying Bob's utterance and the picture) until the participant pressed one of the buttons.

The experiment featured a three-by-three design: crossing Picture (within-participants: Good Control, Bad Control, Target) and QUD (between-participants: wh-word, indefinite, quantifier). Similarly to Experiment 5, there were two types of pictures: Control, of which Bob's sentence was an unambiguously good/bad description, and Target, where the judgment depends on whether the inference has been derived – see Figure 3.2. Bob's answers are good descriptions of the Target pictures on their literal, but not on their inference-enriched reading. Good Control and Target

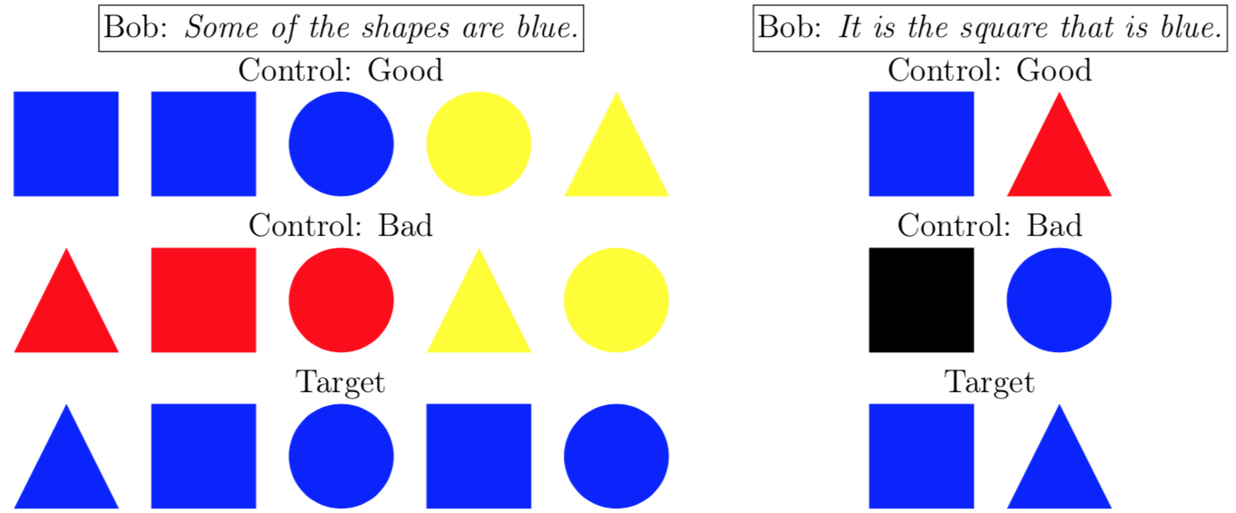

Bob: *Some of the shapes are blue.*
Control: Good
Control: Bad
Target
Bob: *It is the square that is blue.*
Control: Good
Control: Bad
Target

Figure 3.2: Example experimental trials from Experiment 6: SI (left) and EXH (right)

pictures were identical to those used in the elicitation experiment. Note that for Experiment 6, Bad Controls were also added. SI Bad Control pictures contained five shapes, none of which has the color (here, blue) mentioned in the sentence. EXH Bad Controls contained the shape mentioned in the sentence, but in a color different from the one in the sentence (here, black square), while the mentioned color showed up on a different shape (here, blue circle).

The QUD condition (i.e. Anne's questions) was also a three-level manipulation: wh-word, indefinite, and quantifier questions. These questions were the most frequent elicited questions in Experiment 5 (see Table 3.2), with the exception of the *both*-question (in EXH), which I added as a counterpart to the *all*-question (in SI). (12)-(13) provide examples of the question manipulation for each construction. (Wherever there are two questions listed, half the participants saw one variant, and half the other variant.)

(12) QUD manipulation in SI (boldface for illustration only)
wh-word: **What** color are the shapes?
indefinite: Are there **any** blue shapes?/Are **any** shapes blue?
quantifier: Are **all** shapes blue?

(13) QUD manipulation in EXH (boldface for illustration only)
wh-word: **Which/**What shape is blue?
indefinite: Are there **any** blue shapes?
quantifier: Are **both** shapes blue?

QUD was manipulated between participants, placing each participant in either the wh-word, the indefinite, or the quantifier QUD condition. To prevent fatigue effects due to only encountering the same type of question, experimental items were intermixed with filler items, which included different shapes and questions unrelated to either inference, or QUD type. Each participant saw 12 SI, 12 EXH and 12 filler trials. The experiment was administered in a Latin Square design.

3.4.3 Predictions

Existing experimental work has shown that QUDs affect how likely an implicature is to be calculated (Section 3.1.2) – an effect I predict to extend to the previously untested *it*-cleft exhaustivity. Based both on previous empirical results (Zondervan et al., 2008; Degen, 2013) and theoretical proposals (Hulsey et al., 2004), we can make the prediction that for SI, *all*-QUDs would result in the highest rate of implicature calculation. *Any*-QUDs, on the other hand (see e.g. (3)), are predicted to bias against deriving the implicature and lead to lower calculation rates. For EXH, parallel predictions can be made for the corresponding quantifier and indefinite: *both*-QUDs are predicted to lead to higher rates of implicature calculation than *any*-QUDs. The predictions for wh-question QUDs (SI *what*-QUD, EXH *which*-QUD) are less clear, in part because they have previously not been treated as relevant QUDs in these contexts, and it is less obvious what existing theoretical frameworks would predict about them (I discuss this in more detail in Section 3.5). I therefore treat their biasing behavior as an empirical question. When analyzing the results, my primary focus will be on the two other QUDs (*any* and *all* for SI; *any* and *both* for EXH), and my analysis of the wh-questions will be more exploratory.

Implicature calculation rates are indexed by the proportion of ‘Not Good’ responses to Target pictures: if a participant says that Bob gave a ‘Not Good’ description of a Target picture, she has calculated the SI/EXH inference. I thus predict variation across QUDs in the percentage of ‘Not Good’ responses to Target. For example, *any*-QUDs are predicted to result in lower rates of calculation and therefore lower ‘Not Good’ percentages for Target pictures. In what follows, I make a distinction between Literal-biasing QUDs, which lead to lower rates of inference generation (the prediction for e.g. *any*) and Inference-biasing QUDs, which lead to higher rates of inference generation (the prediction for e.g. *all*, *both*).

I consider longer RT when responding ‘Not Good’ to Target, relative to the RT when responding ‘Not Good’ to (Bad) Control, to be what indexes the cost of implicature calculation. This is because Bad Controls can be rejected based on literal, semantic meaning, but the rejection of a Target picture suggests that the participant has gone through the inference calculation process. In conducting such an analysis, I depart from Chemla and Bott (2014) and van Tiel and Schaeken (2017), who additionally analyzed ‘Good’ responses and focused on the interaction of Condition and Response. My reason for doing so is that responding ‘Good’ to Target pictures might indicate implicature non-calculation and reasoning only with the literal meaning, but it is also consistent with participants calculating, and then cancelling the implicature. Thus it is possible that for at least some participants or trials, responding ‘Good’ to Target may have included going through the inference calculation process. This would have led to increased RTs, introducing a confound for the interpretation. For this reason, in my analysis I focus on the Target vs. Control difference when responding ‘Not Good’, and disregard ‘Good’ responses to Targets.

Given the QUD hypothesis (Degen, 2013; Degen and Tanenhaus, 2014), we can predict that under Inference-biasing QUDs (i.e. those that bias towards deriving an implicature) implicature derivation will not incur a cost. That is, we should not see a difference in ‘Not Good’ to Target as compared to ‘Not Good’ to Control RTs in Inference-biasing QUD conditions. On the other hand, Literal-biasing QUDs (i.e. those that bias against deriving an implicature) will have the effect that when the implicature is derived, its calculation is a costly and therefore slower process. Under Literal-biasing QUDs, therefore, we would expect an increase in the time it takes to respond ‘Not Good’ to Target as compared to ‘Not Good’ to Control. On the contrary, instead of a difference in processing cost tracking QUDs, the Lexical Access hypothesis (van Tiel and Schaeken, 2017) predicts a difference in processing cost between SI and EXH —see Table 3.1 for a summary of the diverging predictions of the two accounts. That is, given Lexical Access, increased RTs are predicted for responding ‘Not Good’ to Target as compared to ‘Not Good’ to Control in the case of SI, but no such difference is predicted for EXH.

3.4.4 Results and analysis: rate of inference calculation

Prior to data analysis, I removed trials with extremely short or long response times by excluding the top and bottom 2,5% of the data based on reaction times.

Figures 3.3 and 3.4 plot the proportions of ‘Not Good’ responses for SI and EXH respectively. The primary purpose of Good and Bad Control pictures was to make sure participants are adequately doing the task, which we can see from responses being largely at floor and ceiling, respectively. Therefore, in the following, I focus on the analysis of the more informative Target pictures, which constitute our critical manipulation. Recall that for Target pictures, a 'Not Good' answer indicates implicature calculation. For the statistical analysis, a logistic regression model (glm function in R) was fit (mixed effects models did not converge), predicting Response (Good vs. Not Good) by QUD. Because the main prediction concerns the *any*-*all/both* difference, levels within the QUD variable were treatment coded, with *any* serving as the reference level. In SI, the statistical analysis revealed that the difference between *all* vs. *any* ($p < 0.001$) is significant, while *what* resulted in responses somewhere in between the two other QUDs, with the *any* vs. *what* ($p < 0.06$) difference being only marginally significant (see Figure 3.3 and Table 3.3). An additional pair comparison between *all* and *what* revealed a marginally significant difference ($\beta = -0.58$, $z = -1.93$, $p < 0.06$). In EXH, *any* vs. *both* ($p < 0.001$) was revealed to be significantly different, but the *any* vs. *which* difference is not significant ($p = 0.7$) (see Figure 3.4 and Table 3.4). An additional pair comparison between *both* and *which* revealed a significant difference ($\beta = -1.64$, $z = -4.64$, $p < 0.001$).

	Estimate	Std. Error	*z* value	*p* value
Intercept (any)	-0.06	0.19	-0.29	0.77
all	1.11	0.3	3.75	<0.001
what	0.53	0.28	1.92	0.05

Table 3.3: Results of Experiment 6: *Scalar inference*: Parameter estimates, standard errors, *z* values and *p* values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD.

	Estimate	Std. Error	*z* value	*p* value
Intercept (any)	0.33	0.19	1.71	0.09
both	1.55	0.36	4.36	<0.001
which	-0.09	0.27	-0.35	0.73

Table 3.4: *Results of Experiment 6: It-cleft exhaustivity*: Parameter estimates, standard errors, *z* values and *p* values from a logistic regression model of the 'Not Good' vs. 'Good' responses to Target, predicted by QUD.

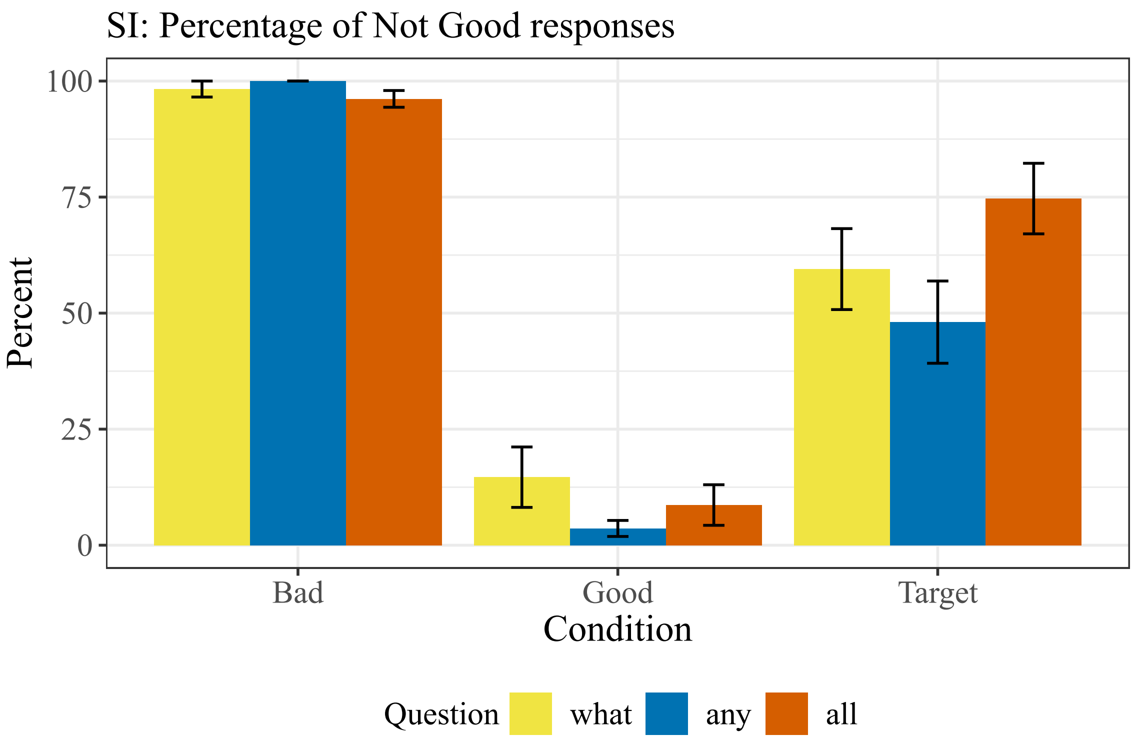


Figure 3.3: Results of Experiment 6: Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *scalar inference*. Different colors denote the different QUD conditions. Error bars represent standard error.

My prediction was that QUDs would modulate the rate of deriving pragmatic inferences, as indexed by rates of judging Bob's answer (as a description of the Target picture) 'Good' vs. 'Not Good'. The findings are in line with these predictions. For SI, I found significantly fewer implicatures calculated under *any*-QUDs than under *all*-QUDs: that is, *any* is a Literal-biasing, while *all* is an Inference-biasing QUD. *What*-QUDs fall in the middle, making it unclear whether they can be categorized as either Literal- or Inference-biasing. In EXH, I also found significant differences in the rate of implicature calculation, successfully extending earlier findings to a different kind of inference. Specifically, significantly fewer implicatures were calculated under *any*- and *which*-QUDs than under *both*-QUDs. In other words, for EXH, *any* and *which* are Literal-biasing, while *both* is an Inference-biasing QUD.

It is worth noting that under *any*-QUDs, there is a presuppositional mismatch between the EXH target sentence and the question: the EXH construction carries the existential presupposition that

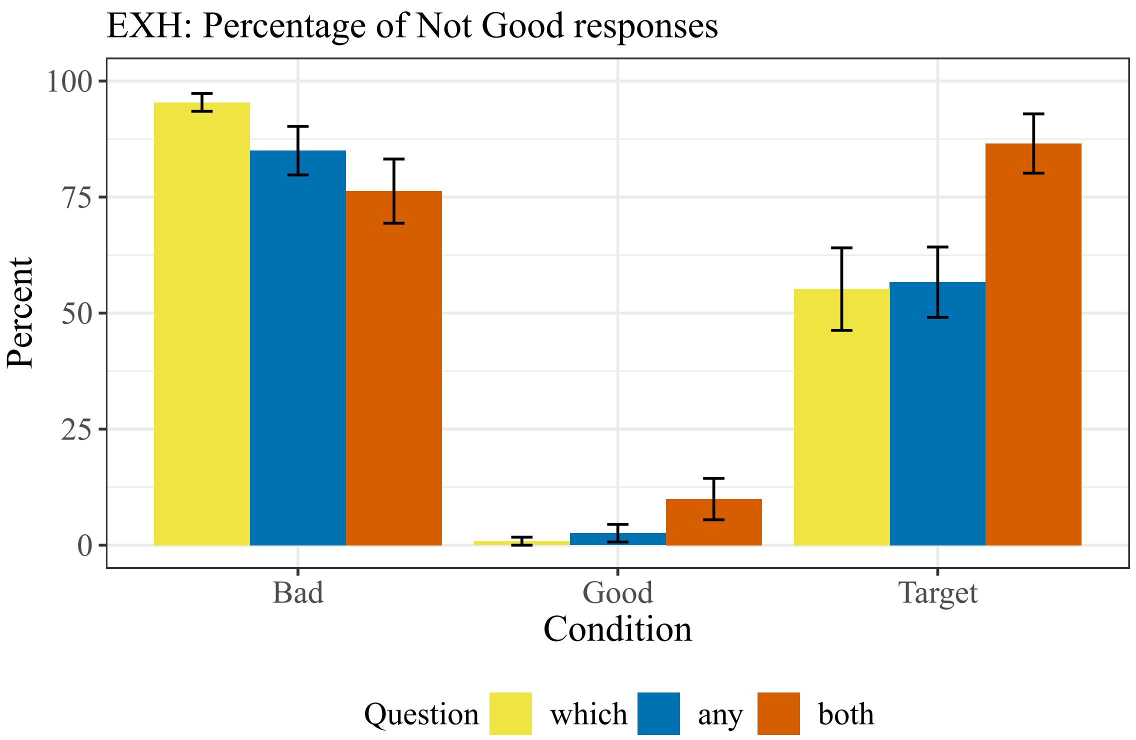


Figure 3.4: Results of Experiment 6: Proportion of participants' 'Not Good' (as opposed to 'Good') responses by Picture condition for *it-cleft exhaustivity*. Different colors denote the different QUD conditions. Error bars represent standard error.

something is blue, while *any*-questions are not presuppositional. But it is instructive to note that in the Good Control condition, the rate of 'Good' responses was at ceiling. That is, participants almost always deemed the EXH sentence a 'Good' answer to an *any*-question, given a picture compatible with the exhaustivity inference. This suggests that participants were able to accommodate the existential presupposition and generate the exhaustivity inference, which is the phenomenon of interest to us.

Now that we have established that for both SI and EXH, some QUDs bias towards deriving the inference, while other QUDs bias against it, I turn to my main hypothesis. I show that these differences among QUDs are reflected in reaction time cost: the implicature calculation process will incur more or less reaction time cost depending on whether the target sentence is in a supportive context (i.e. under Inference-biasing QUDs) or a non-supportive context (i.e. under Literal-biasing QUDs). I analyze reaction time results in the next section.

3.4.5 Results and analysis: reaction times

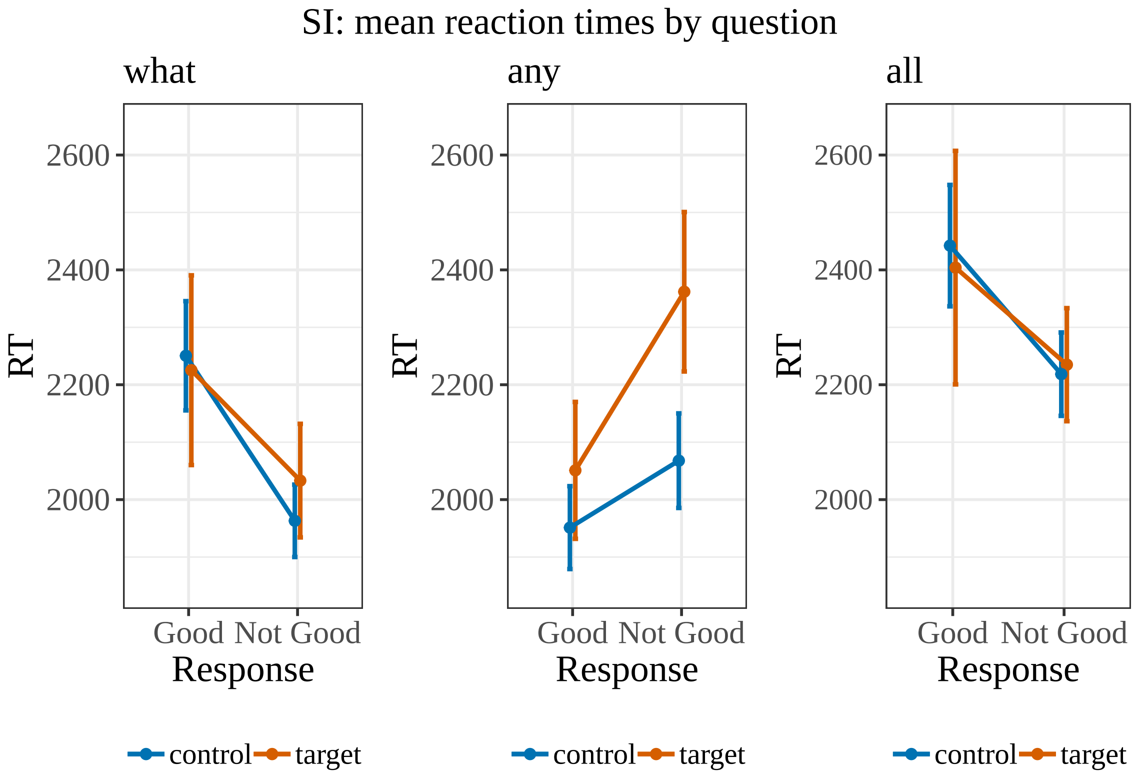


Figure 3.5: Results of Experiment 6: Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *scalar inference*. Different QUD conditions are displayed on separate plots. Colors denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error.

Figures 3.5 and 3.6 plot reaction times broken down by QUD. Because my predictions concern the difference in RT when responding 'Not Good' to Target, as compared to Control (see Section 3.4.3), and I do not have specific hypotheses about differences in RTs when responding 'Good', I restrict the statistical analysis to 'Not Good' responses. Nevertheless, the full data set is plotted in Figures 3.5-3.6. For the statistical analysis, a linear mixed effects model (lmer from the lme4 package in R, Bates et al., 2015) was fit, predicting RT by Condition (Target vs. Control). Levels within Condition were treatment coded, with Control serving as the reference level. Random slopes and intercepts were included for participants and items. Whenever the full model did not converge, the random effects structure was simplified following the recommendations of Barr et al. (2013).

The *p* values reported below were estimated using the Satterthwaite procedure, as implemented in the lmerTest package in R (Kuznetsova et al., 2017). In the following I analyze RT data question-by-question, following the predictions made under the QUD hypothesis.

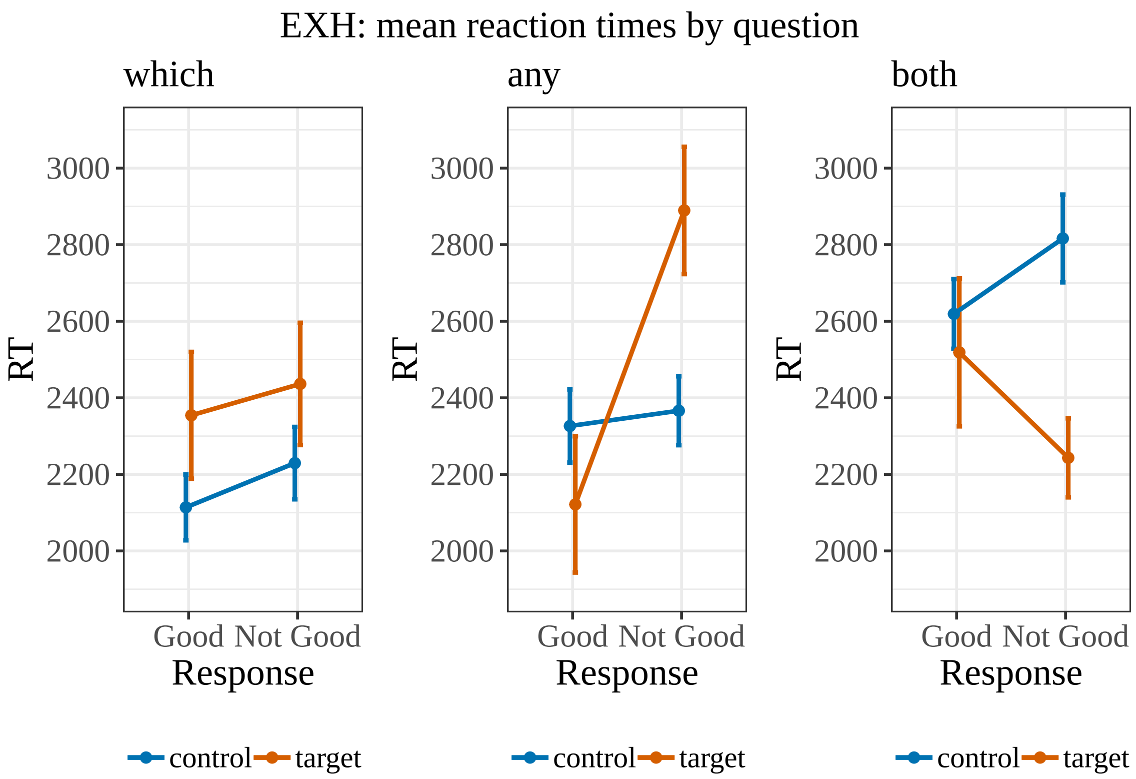


Figure 3.6: Results of Experiment 6: Mean reaction times (RT) in ms for judging Bob's answer as 'Good' or 'Not Good' in *it-cleft exhaustivity*. Different QUD conditions are displayed on separate plots. Colors denote Control (Good and Bad) vs. Target pictures. Error bars represent standard error.

Any-QUDs

Recall that *any*-QUDs were predicted to bias toward the literal meaning based on earlier theoretical and experimental work, and that is indeed what we found in the inference calculation rate data. Thus *any*-QUDs are predicted to make calculating an SI/EXH inference a costly process, manifested in longer RTs when responding 'Not Good' to Target as compared to Control. For the SI *any*-QUD, I found a significant difference in RT between responding 'Not Good' to Target vs. Control ($\beta = 391.47$, $t = 2.6$, $p < 0.05$). Similarly, for the EXH *any*-QUD, I found a significant difference in RT between responding 'Not Good' to Target vs. Control ($\beta = 539.01$, $t = 3.3$, $p < 0.01$). Importantly, both patterns (see middle panels in Figures 3.5 and 3.6) show increased RT for the Target picture, as compared to the Control picture when responding 'Not Good'. This is thus in line with the prediction that *any*-QUDs make inference computation time-consuming.

All- and both-QUDs

The predictions made for the *all*- and *both*-QUD are the opposite of the prediction about the *any*-QUDs. *All*- and *both*-QUDs are Inference-biasing, and should therefore not lead to increased RTs when responding 'Not Good' to Target. For the SI *all*-QUD, there was no significant difference in RT between responding 'Not Good' to Target vs. Control ($\beta = 5.06$, $t = 0.04$, $p = 0.97$). For the EXH *both*-QUD, I found a significant difference in RT between responding 'Not Good' to the two types of pictures ($\beta = -478.01$, $t = -2.82$, $p < 0.01$), such that responding to Control resulted in longer RTs than responding to Target. That is (see the rightmost panel in Figure 3.6), while there was no reaction time cost for calculating the inference (i.e. no increased RTs for Target), I found an unexpected cost for responding 'Not Good' to the Control picture.

I argue that this unexpected cost is a side-effect of the picture stimuli. In particular, the verification process involved in rejecting a Bad Control involves two steps for EXH, but not for SI. For EXH (see Figure 3.7), given Bob's utterance *It is the square that is blue*, both the color and identity of each shape needs to be checked. This is because there is in fact only one blue shape (cf. exhaustive meaning) in the Bad Control picture, but that blue shape is not the correct one (i.e. the square). Note that all EXH stimuli necessarily have to include a blue shape, because the *it*-cleft sentence also carries an existential presupposition that something is blue. In contrast, for SI (see Figure 3.7), Bob's utterance *Some of the shapes are blue* can be rejected in one step, because no shapes in that display are blue.

Nevertheless, the findings are generally in line with the predictions of the QUD hypothesis in that the Inference-biasing *all*- and *both*-QUDs did not lead to a delay in reaction times for deriving

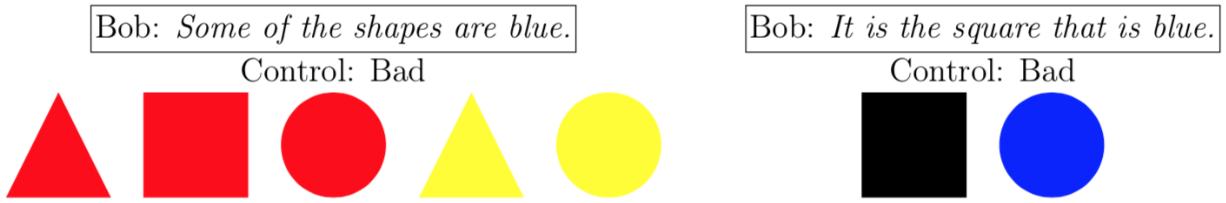

Bob: *Some of the shapes are blue.*
Control: Bad
Bob: *It is the square that is blue.*
Control: Bad

Figure 3.7: Example experimental trials from Experiment 6: SI (left) and EXH (right)

the SI/EXH inference.

Based on the above, we see a difference between *any*-QUDs, which led to processing cost, and *all*- and *both*-QUDs, which did not. To confirm that RT patterns vary across QUDs, I conducted an additional analysis focusing on the interaction of Condition with QUD. A linear mixed effects model (lmer) was fit, predicting RT by Condition (Target vs. Control), QUD (SI: *any* vs. *all*; EXH: *any* vs. *both*) and their interaction. Both variables were sum-coded, and models included random effects as described at the beginning of Section 3.4.5. For SI, I found a significant interaction of Condition with QUD ($\beta = 96$, $t = 2.1$, $p < 0.05$). Similarly, for EXH, I found a significant interaction of Condition with QUD ($\beta = 251.21$, $t = 4.62$, $p < 0.001$). That is, for both SI and EXH sentences, the RT patterns signalling processing cost were found to vary according to the preceding QUD.

What- and which-QUDs

Recall that we were not able to make clear predictions about *what*- and *which*-QUDs (Section 3.4.3), despite their frequency in the elicitation experiment. For the SI *what*-QUD, I found no significant difference in RT between responding ‘Not Good’ to Target vs. Control ($\beta = 194.91$, $t = 1.74$, $p = 0.1$). That is, the *what*-QUD parallels the *all*-QUD in resulting in no difference in RTs when responding ‘Not Good’ to Control vs. Target. This suggests that despite the mixed results in terms of SI calculation rate, the *what*-QUD is Inference-biasing based on reaction time measures. As for the EXH *which*-QUD, there was no significant difference in RT between responding ‘Not Good’ to Target vs. Control ($\beta = 266.6$, $t = 1.61$, $p = 0.11$). Qualitatively the *which*-QUD shows a somewhat similar pattern to the *any*-QUDs, but the results are much less clear.

In sum, the predictions of the QUD hypothesis are largely borne out in the data: the questions which could be unambiguously classified as Literal- (*any*) vs. Inference-biasing (*all, both*) show divergent behavior in terms of reaction time cost. *Any*-QUDs resulted in SI and EXH calculation being time-consuming, while the SI *all*-QUD and EXH *both*-QUD facilitated reaction times. The SI *what*-QUD patterned with Inference-biasing QUDs in that it did not lead to increased reaction times, while the EXH *which*-QUD was qualitatively similar to Literal-biasing *any*-QUDs. Importantly, the results do not reveal that SI calculation always leads to processing cost, but EXH calculation does not, which would have been predicted by the Lexical Access hypothesis. I discuss the implications of these results in the next section, along with the calculation rate results.

3.5 General discussion

In this chapter I tested the hypothesis that the likelihood of calculating SI and EXH pragmatic inferences, as well as the processing cost of that calculation, tracks the QUD – a prediction of constraint-based models of implicature calculation. Under such a model, the more probabilistic support there is from multiple cues, the more quickly and robustly listeners will compute pragmatic inferences, with the QUD being one of the relevant cues. I elicited explicit questions to approximate the relevant QUDs (Experiment 5). Using these QUDs in Experiment 6, I found that they fall into one of two classes: under Literal-biasing QUDs, the rate at which participants drew inferences was lower, and under Inference-biasing QUDs, the rate at which they drew inferences was higher. Differences among QUDs also predicted processing cost. Under the Literal-biasing QUDs, making an inference-enriched judgment took longer than responding to the relevant literal control, whereas a facilitation of reaction time for such inferences was observed under Inference-biasing QUDs. By and large, these patterns hold for the majority of the questions examined, although it is also worth noting that there appeared to be more nuanced patterns with wh-questions (see more discussion below). Most crucially, we did not observe across-the-board processing cost for deriving implicatures, nor did we observe that computing inferences is always cost-free. Instead, the results strongly suggest that the cost of computing SI and EXH inferences is context-dependent – it is costly when the target expression was preceded by non-supportive QUDs. Such context-dependent cost imposes a significant constraint on our hypothesis space. For example, these results would be challenging for the *default hypothesis* and the *literal-first hypothesis* introduced in Section 3.1.1, since both would predict categorical behavior that is QUD-independent.

The current findings also pose an empirical challenge for the Lexical Access account. While evidence for Lexical Access comes from studies that presented target sentences in the absence of any context, Experiment 6 varied the QUD. My results do not support the prediction that the calculation of SI inferences, but not the calculation of EXH inferences, would lead to processing cost. Instead, I found that depending on context, both inferences can lead to a reaction time delay (or the lack of a delay). This effect of context-dependence is unexpected if the cost of calculating pragmatic inferences is directly and uniquely tied to the construction and complexity of the relevant alternatives. Particularly informative in this respect are the findings about EXH, as well as SI *what*-QUDs. The predictions of Lexical Access converge with that of a QUD-based hypothesis about processing cost in the case of questions that explicitly mention alternatives, e.g. *all*-QUDs. Following a context (in Experiment 6: an explicit question) where the alternative *all* has been made salient, lexical retrieval of *all* will likely be a faster process. However, it is less clear how the Lexical Access hypothesis would capture the findings about questions that do not explicitly mention relevant lexical alternatives: the *what*-QUD, which showed reaction time patterns similar to the *all*-QUD, or the findings about EXH, where no lexical alternatives are relevant. It is possible that both lexical retrieval and context-dependence contribute to the complexity of generating pragmatic inferences. My results do not necessarily rule out that lexical access plays a role. Future research should probe further whether these diverse sources of complexity could selectively target different aspects of the processing cost generated during the inferential process.

The findings reported here contribute to the empirical landscape in a number of additional ways. First, I probed the robustness of earlier work on the QUD-sensitivity of scalar inference calculation and processing. Using a different experimental paradigm (sentence-picture verification) than existing work on QUDs, my findings successfully replicate the results of i.a. Zondervan et al.

(2008), Degen (2013) and Cummins and Rohde (2015) (for calculation rates) and are in line with trends observed by Degen (2013) and Degen and Tanenhaus (2014) (for reaction time). Moreover, this was done by manipulating the QUD directly via explicit questions, rather than implicitly via background stories – which the earlier processing experiments utilized. Crucially, I also found that QUD-sensitivity extends beyond the well-known case of scalar inference, to a previously untested pragmatic inference: *it*-cleft exhaustivity. Lastly, Experiment 5 took an important step towards better understanding how to empirically probe relevant QUDs, which I return to at the end of this section.

The specific role of QUDs observed in this chapter could be formalized under the Question-Answer Requirement (QAR, Hulsey et al., 2004) approach:

(14) *The Question Answer Requirement (QAR)*

The selected interpretation of an ambiguous sentence, whether true of false, is required to be a good answer to the Question Under Discussion. (A good answer is an interpretation that at least entails an answer to the QUD.)

In other words, any sentence is to be understood as an answer to a question, this question being the QUD (Hulsey et al., 2004; Gualmini et al., 2008). The QAR posits that the selected interpretation of an ambiguous sentence is required to be a good answer to the QUD. The two interpretations that SI target sentences allow for are repeated in (15) below.

(15) Some of the shapes are blue.

Literal: At least some of the shapes are blue.

Inference-enriched: Some but not all of the shapes are blue.

Using a standard Hamblin semantics for questions (Hamblin, 1976), the SI results can be accommodated under the QAR as follows. On such a semantics, the meaning of a question, including the meaning of a QUD, is a partition of the set of possible worlds. The meaning of the QUD *Are* *there any blue shapes?* is a partition of worlds into the two sets (16-a) and (16-b):

(16) a. $\{w : \exists x.x \text{ is a blue shape in } w\}$

(*At least some shapes are blue.*)

b. $\{w : \neg\exists x.x \text{ is a blue shape in } w\}$

(*No shapes are blue.*)

Here, it is not necessary to derive the SI inference from *Some of the shapes are blue* in order to provide a good answer in the sense of the QAR, because the literal interpretation *At least some of the shapes are blue* corresponds to (16-a). Deriving the inference also results in a good answer, because the inference-enriched interpretation (*Some but not all of the shapes are blue*) entails (16-a). Therefore, under the *any*-QUD, the target sentence is a good response whether or not the inference is derived.

However, the picture changes for the *some but not all* SI target sentence under the QUD *Are all shapes blue?*, which partitions the set of worlds into the following two sets (assuming a five-shape display):

(17) a. $\{w : \forall x.x \text{ is a shape} \rightarrow x \text{ is blue in } w\}$

(*All shapes are blue.*)

b. $\{w : \neg\forall x.x \text{ is a shape} \rightarrow x \text{ is blue in } w\}$

(*Not all shapes are blue.*)

In this case, only if the inference is derived is the dialogue QAR-compliant, because the literal reading does not address the QUD. However, the enriched reading *Some but not all of the shapes are blue* entails (17-b). Thus, the SI target sentence is only a good response to the *all*-question on the inference-enriched reading. This is reflected in the finding that participants derived significantly more implicatures with the *all*- than with the *any*-QUD, and that SI calculation led to a reaction time cost under the *any*-, but not under the *all*-QUD.

As for EXH, the two potential interpretation of the target sentence are repeated in (18).

(18) It is the square that is blue.

Literal: The square is blue.

Inference-enriched: Only the square is blue.

Assuming a two-shape display, the QAR captures the EXH findings in a way parallel to the SI findings. The EXH *any*- and *both*-QUDs result in the same partitioning of the set of worlds as the SI *any*- and *all*-QUDs respectively. The only difference is that the domain of quantification is now two instead of five, given the experimental pictures – as is reflected in the English paraphrases below.

(19) Partitioning from *any*-QUD (*Are there any blue shapes?*):

a. $\{w : \exists x.x \text{ is a blue shape in } w\}$

(*At least one shape is blue.*)

b. $\{w : \neg\exists x.x \text{ is a blue shape in } w\}$

(*Neither shape is blue.*)

(20) Partitioning from *both*-QUD (*Are both shapes blue?*):

a. $\{w : \forall x.x \text{ is a shape} \rightarrow x \text{ is blue in } w\}$

(*Both shapes are blue.*)

b. $\{w : \neg\forall x.x \text{ is a shape} \rightarrow x \text{ is blue in } w\}$

(*It is not the case that both shapes are blue.*)

Under the *any*-QUD, both the literal (*The square is blue)* and the inference-enriched (*Only the square is blue*) readings entail (19-a). Therefore, both constitute good answers according to the QAR. However, under the *both*-QUD, only the inference-enriched reading answers the question by entailing (20-b); the literal reading does not bear on the QUD. This difference is reflected in the empirical data: significantly more implicatures were derived under the *both*- than under the *any*-QUD, and only under the *any*-QUD was the computation time-consuming.

Additional to the main finding that context modulates the calculation and processing of SI and EXH, there remain some empirical puzzles regarding the interplay of production and interpretation of QUDs. In Experiment 5, I took a first step in addressing the problem of narrowing down potential QUDs for a given context and conducted an elicitation study, the results of which fed into the QUD manipulation experiment. In doing so, I went beyond previous work that relied only on theoretically informed introspection to identify what may serve as a relevant QUD, and instead I treated this issue as an empirical question. Based on the results of the elicitation experiment, I focused on three types of questions: *any*-, *all/both*- and wh-QUDs. While the quantifier-QUDs have been discussed in existing literature as being relevant QUDs to the SI and EXH target sentences, the wh-QUDs (SI *what*-QUD, EXH *which*-QUD) have not. Yet it is interesting to observe that the novel wh-QUDs were the ones that proved most frequent in the elicitation experiment.

In addition to their novelty, the wh-QUDs also have some other puzzling properties. For instance, the focus structures of the *what*-QUD and the SI answer are potentially incongruent. The *what*-QUD focuses colors, yielding a set of alternatives of the form {*The shapes are red, The shapes are blue,* ...}; but one might think that the inference from the SI target sentence is most natural with focus on the quantifier *some* in *Some of the shapes are blue*, with an attendant set of focus alternatives {*None of the shapes are blue, Some of the shapes are blue, All of the shapes are blue*}. The wh-QUDs also showed more nuanced results in Experiment 6. The SI *what*-QUD did not fall neatly into the Literal- or Inference-biasing category based on the rate of inference calculation, though in RT results it qualitatively patterned with the SI *all*-QUD. As for the EXH *which*-QUD, while it showed distinct Inference-biasing behavior in terms of likelihood of inference calculation, it resulted in mixed RT patterns. On the other hand, the QUDs that showed the clearest patterns (*any*-, *all*-QUDs) are in fact the ones with the closest link to existing theoretical work. Future work should thus address this tension between the observed mixed biasing and processing behavior of *what*- and *which*-QUDs, and their apparent popularity with naive participants in an elicitation task.

The fact that elicitation resulted in somewhat surprising QUDs raises the issue of how to successfully probe theoretical constructs in an empirically grounded manner. One thing lacking in the current empirical picture is a measure of whether participants calculated the SI and EXH inferences in the elicitation experiment. Gathering such a measure could help address open questions regarding why the Target and Good Control conditions elicited very similar questions, as well as the surprising finding that wh-questions, which are not fully congruent with the target sentences in their presuppositions or focus structure, were also elicited. Conducting an elicitation experiment that also targets interpretation would thus be a valuable avenue for future work. Experiment 5 constitutes only a first step in understanding how QUDs can be elicited experimentally.

3.6 Summary of findings

The processing of implicatures has been a central question in linguistics, as it serves as a window into the integration of semantic and pragmatic knowledge and may help answer the question: is one kind of information privileged over the other in reasoning, meaning calculation, and (real-time) processing? A large body of existing work has found that generating scalar inferences incurs a processing cost, evidenced by e.g. increased reaction times. This chapter compared two different hypotheses about what aspect of pragmatic inference calculation causes the observed processing cost. The Lexical Access hypothesis (van Tiel and Schaeken, 2017; Chemla and Bott, 2014) links the processing cost of inference calculation to the retrieval of lexical alternatives, while the QUD hypothesis (Degen, 2013; Degen and Tanenhaus, 2014) proposes that congruence with the QUD is what modulates the processing cost. I compared scalar inference and *it*-cleft exhaustivity, and showed that for both, the likelihood of inference calculation, as well as the attendant processing cost, are better explained by properties of the discourse context than by whether alternative-retrieval is involved.

CHAPTER 4

FACTORS EXPLAINING SCALAR DIVERSITY

As introduced in Chapter 1 (Introduction), scalar inferences like *some but not all* are taken to arise via hearers' reasoning about stronger alternatives (like *all*), upon encountering an utterance containing a weaker scalar term (like *some*)[1]. As we saw in Chapter 2, such stronger alternatives are not merely useful for a theoretical account of how scalar inference arises, but are in fact also retrieved and activated in the real-time processing of scalar inference-triggering sentences. Though, as we saw in Chapter 3, the retrieval of alternatives is not what explains the processing cost associated with scalar inference calculation.

In the next two chapters I turn to a recent puzzle for scalar inference, called *scalar diversity*. Scalar diversity is the observation that the likelihood of inference calculation differs robustly across lexical scales, e.g, *some but not all* is much more likely to arise than the parallel inference *good but not excellent*. This presents a complication for theory in that if scalar inference is 'simply' a matter of reasoning about a stronger alternative to what was said, then we do not expect any variation in how likely this reasoning is to go through, and how likely scalar inferences are to arise. One way to reconcile the core idea of alternative-based reasoning with the puzzle of scalar diversity is to derive the likelihood of scalar inference calculation from properties of the stronger alternatives. More concretely, to identify some property of the alternatives *all* vs. *excellent* that could explain why the former leads to scalar inference more robustly than the latter, even though they are both a stronger alternative to *some* and *good* respectively.

In this chapter, I first introduce existing work on scalar diversity (Section 4.1). Then I report on corpus work that provided a new set of lexical scales, which is larger and better balanced across grammatical categories than those used in prior literature (Section 4.2). I then replicate the scalar diversity phenomenon on this new set of scales, and propose an information theoretic measure to better quantify the observed variation than has been done in prior work (Section 4.3). The majority

[1] Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/n4d6t/?view_only=07799f6001f54b31a6b088720278899d

of this chapter is then concerned with identifying and testing various factors that might explain scalar diversity —most, but not all, of which have to do with properties of alternatives (Section 4.4).

4.1 Scalar diversity

Scale	Sources
<*some, all*>	Noveck (2001); Noveck and Posada (2003); Papafragou and Musolino (2003); Bott and Noveck (2004); Feeney et al. (2004); Guasti et al. (2005); Breheny et al. (2006); De Neys and Schaeken (2007); Pouscoulous et al. (2007); Banga et al. (2009); Geurts and Pouscoulous (2009); Huang and Snedeker (2009); Clifton and Dube (2010); Grodner et al. (2010); Barner et al. (2011); Chemla and Spector (2011); Bott et al. (2012); Geurts and van Tiel (2013); van Tiel (2013); Degen and Tanenhaus (2014)
<*or, and*>	Noveck et al. (2002); Storto and Tanenhaus (2005); Breheny et al. (2006); Chevallier et al. (2008); Pijnacker et al. (2009); Zondervan (2010); Chemla and Spector (2011)
<*might, must*>	Noveck (2001)
<*start, finish*>	Papafragou and Musolino (2003)

Table 4.1: Scalar expressions used in a representative sample of experiments on the interpretation, development and processing of scalar inferences. Table 2 from van Tiel et al. (2016, p. 139).

Much of the initial research on the processing (or acquisition) of scalar inferences has concentrated on two lexical scales: <*some, all*> and <*or, and*>. Table 4.1, repeated from van Tiel et al. (2016), provides an overview of the scales that have been used in an illustrative sample of experimental research on scalar inference. The tacit assumption in this body of work is that these scales are representative of the entire family of scalar expressions, and experimental findings would generalize to all scalar inferences. Take, for instance, the example in (1), which is based on the <*good, excellent*> scale.

(1) The movie is good.

Literal meaning: The movie is at least good.

Inference-enriched meaning: The movie is good, but not excellent.

Such examples, in principle, give rise to SI the same way as <*some, all*>, e.g. comprehenders, upon encountering *The movie is good*, will reason about *The movie is excellent* as a potential alternative the speaker could have uttered and infer its negation. More generally, because all scalar inferences are assumed to be derived by the hearer reasoning that an informationally stronger alternative is not true (because the speaker chose not to utter it), there should be no difference across scales in the robustness of calculation, processing, or acquisition. However, this assumption, referred to as the uniformity assumption (van Tiel et al., 2016, p. 139), has been challenged by more recent work, which found that there is in fact considerable variation across different scales in the rates of SI calculation.

4.1.1 Early evidence against uniformity

Despite a large majority of studies on scalar inference concentrating on a very small number of scales, making the tacit assumption of uniformity, these studies already contained some indication that scales are not uniform in how likely they are to lead to SI. Van Tiel et al. (2016) reviews a number of studies that provide (in their words) "extant evidence for diversity" (p. 140). Geurts (2010), for instance, observed that across the experiments he surveyed, the mean rate of SI calculation was higher for <*some, all*> than for <*or, and*>. In the developmental literature there exists also early evidence that scales show non-uniform behavior. Papafragou and Musolino (2003) compared <*some, all*>, <*two, three*> and <*start, finish*> and found different rates of SI calculation in children, with rates for the numeral being higher. Barner et al. (2011) similarly found that children draw SIs at different rates between <*some, all*> and ad hoc scales, with ad hoc SI rates being higher. It must be noted, however, that the main purpose of these studies was not to test the uniformity assumption.

One of the first studies to directly compare different (classes of) scales was Beltrama and Xiang (2013), who tested adjectives (e.g. <*decent, good, excellent*>) and modals (e.g. <*possible, likely, certain*>). This study employed a self-paced reading task, with acceptability ratings about plausibility. An example of the stimuli used can be seen in (2).

(2) a. Mark is a decent/good/excellent student. That's why he has been accepted to Harvard for a Ph.D.

b. Sofia is a decent student. That's why it's possible/likely/certain that she will get into Harvard.

Beltrama and Xiang (2013) found that in adjectival scales, the lowest member (*decent*) of the scale triggered SI, as evidenced by low plausibility ratings, the logic of this experimental design being that if Mark is a *decent but not excellent* student, it should not be plausible that he was accepted to Harvard. The middle member of the adjectival scales (*good*), however, received higher ratings, suggesting that it did not trigger the *not excellent* SI. Importantly, Beltrama and Xiang (2013) did not find a significant difference between low and middle members in the case of modal scales. This therefore constitutes evidence that adjectival and modal scales do not have uniform behavior in triggering SI.

The research group of Doran et al. (2012) and Baker et al. (2009) were also among the first to conduct a systematic experimental testing of the uniformity assumption. Among others, Doran et al. (2012) compared quantifier scales such as <*some, much, all*> to adjectival scales such as <*poor, comfortable, wealthy*>. (Modals, cardinals, rankings, and manner implicatures were also tested, but are omitted from the discussion here). The below example (3) illustrates their experimental task.

(3) a. Irene: How much cake did Gus eat at his sister's birthday party?
Sam: He ate most of it.
FACT: By himself, Gus ate his sister's entire birthday cake.

b. Irene: How would you say Alex is doing financially?
Sam: He's comfortable.
FACT: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.

In this study, participants had to decide whether Sam's answers were true or false. In this task, if a participant deems Sam's statement to be false, that is taken as evidence that they have derived the SI: in (3), from the weaker scalar term *most* or *comfortable*. Doran et al. found that quantified statements were rejected roughly twice as often as sentences with adjectives—in other words, SI calculation was twice as frequent for quantifiers as for adjectives.

4.1.2 Previous studies on scalar diversity

The first large-scale study on scalar diversity was conducted by van Tiel et al. (2016), who focused not only on categorical differences between types of scales based on different parts of speech (e.g. modals vs. adjectives, or quantifiers vs. adjectives), but also on differences across specific lexical scales even when they belong to the same category. The authors also tested an even greater variety of scales (43 in total): quantifiers (e.g. <*some, all*>), adverbs (<*sometimes, always*>), auxiliary verbs (<*may, have to*>), main verbs (<*participate, win*>) and adjectives (<*content, happy*>). They employed an inference task:

(4) John says: *She is intelligent.*
Would you conclude from this that, according to John, she is not brilliant?

Experiment 1 used materials that were as neutral as possible (4), while Experiment 2 used more specific predicates and full noun phrases, e.g. *This student is intelligent*. Participants had to respond by clicking "Yes" or "No", with a "Yes" response indicating SI calculation for the given scale. Overall, van Tiel et al. (2016)'s study found considerable variation in the rates of scalar inference derivation, ranging from 4% (for seven scales) to 100% (for two scales). While there were some minor differences across the two experiments when it comes to the SI rates yielded by specific scales, the overall scalar diversity effect was robust in both of them. This constitutes convincing evidence that the uniformity assumption does not hold: different scales do not lead to scalar inference derivation to the same extent.

In some more recent work, Simons and Warren (2018) found significant differences among different scales even when placing the relevant sentences in a rich context, providing participants with a story. For instance, the passage in (5) was used to probe SI calculation from the scalar terms *good*, *cool*, and *possible*.

(5) Sally went to the pool around 4 o'clock. She enjoyed swimming at the end of the day: she was a good swimmer and she loved how the swim left her feeling cool and refreshed. And although she wouldn't have admitted it to anyone, she went to the pool in part because it was possible she would run into Steven there.

Additionally, this study probed SI calculation without providing participants with explicit scalar alternatives: e.g., the <*many, all*> scale was tested without showing participants the word *all*. Instead, after seeing a sentence like *She noticed that many of her pencils were chewed on*, participants had to judge whether *100% of her pencils were chewed on*. Even though this study used richer and therefore more naturalistic stimuli than previous studies, the pattern of SI rates was similar to van Tiel et al. (2016) (albeit with only 9 scales tested). For instance, the weaker scalar *good* (SI: not rated 10 on a quality scale 1-10) and *think* (SI: not 100% confident/certain) were almost 30% less likely to lead to SI calculation than *many* (SI: not 100%) and *possible* (SI: not 100% chance/likelihood).

Scalar diversity has also recently been shown in ignorance inferences, which are closely related to SI. Alexandropoulou (2022) presented participants with stimuli such as (6):

(6) Maria says: "Kostas' overall performance at school is at least good."
Conclusion: Maria doesn't know whether Kostas' overall performance at school is excellent.

Participants had to rate (on a 1-7 scale) the validity of this conclusion, which tested not SI rate (*good* → *not excellent*) as all studies discussed previously, but rather speaker ignorance (*good* → speaker doesn't know whether *excellent*). The study found that adjectival scales significantly differed from numerals in how strongly they license an ignorance inference: ignorance inferences were less likely to arise with *at least* as an adjectival modifier than as a numeral modifier. All of the above findings thus attest to the pervasiveness of the phenomenon that different scales do not behave uniformly when it comes to SI calculation.

The question arises, then, how to capture this observed variation in SI rates: can we identify some properties of different scales that influence how robustly they lead to SI calculation? Existing work has identified a number of such properties. Van Tiel et al. (2016) hypothesized that the distinctness of lexical scales might explain scalar diversity. Distinctness refers to whether the speaker considered the distinction between the weaker (*some*) and stronger (*all*) scalar terms substantial enough that she would have used the stronger one if possible. Distinctness was indeed found to be a predictor of SI rates. Van Tiel et al. (2016) operationalized distinctness as semantic distance and boundedness, two notions that go back to Horn (1972, p. 112). Measuring semantic distance via a rating task, it was revealed that the more distant a weak and a strong scalar term are, the stronger the SI from the weak term is. This can be intuitively seen on the <*some, many, most, all*> scale: an utterance of *Mary ate some of the deep dish* most strongly implicates that Mary didn't eat all of the deep dish, while the inference *Mary didn't eat most of the deep dish* is less likely, and *Mary didn't eat much of the deep dish* is least likely. The second component of distinctness is boundedness: unbounded scales (e.g., <*good, excellent*>), in which both the weaker and stronger term denote intervals, were found to lead to significantly fewer SIs than bounded scales, in which the stronger scalar term denotes a fixed point or endpoint (e.g., <*some, all*>).

Subsequent work has identified further properties of scales that predict how likely they are to lead to SI. Investigating adjectival scales, Gotzner et al. (2018) found that certain semantic properties of adjectives, such as polarity and extremeness, are relevant for SI calculation. In particular, their results revealed that negative scales (e.g., <*bad, awful*>) yield higher SI rates than positive ones (e.g., <*good, great*>). Additionally, scales in which the stronger term is an extreme adjective (e.g., *excellent* or *huge*) were found to lead to lower SI rates —for findings regarding extremeness, see also Beltrama and Xiang (2013). Existing work has also related scalar diversity to other semantic-pragmatic processes. Sun et al. (2018) investigated propensity for local enrichment, indexed by the naturalness of sentences such as *Mary ate all, so not some, of the deep dish*. This factor was positively correlated with SI rates: as the authors argue, in order for a sentence such as *Mary ate all, so not some, of the deep dish* to be natural and not contradictory, *some* has to be locally be interpreted on its SI-enriched meaning (*some but not all*). Lastly, Gotzner et al. (2018) also showed that SI rates are negatively correlated with the degree of negative strengthening of the stronger scalar term. Negative strengthening is the phenomenon whereby *John is not brilliant* is interpreted as conveying that John is rather stupid, which can be analyzed as a manner implicature (Horn, 1989). In Gotzner et al.'s study, participants saw sentences such as *He is not brilliant*, and were asked whether they can conclude from this statement *He is not intelligent*. Endorsements of this conclusion were negatively correlated with SI rates, suggesting that, at least for some scales, scalar and manner implicatures might stand in competition (Levinson, 2000).

The observed variation in SI rates has been also related to properties of the context, broadly construed. Pankratz and van Tiel (2021) offer a usage-based explanation of scalar diversity, and show that it is predicted by the relevance of the SI at hand. Specifically, they developed a corpus-based measure of relevance, whereby the more relevant an SI is, the more likely it is to occur in so-called scalar constructions (e.g., *It's good but not excellent*) in a corpus. Ronai and Xiang (2021) investigated the role of the Question Under Discussion in explaining scalar diversity. They hypothesized that, given that experiments typically present SI-triggering sentences in the absence of any context, variation across scales in what implicit discourse context they bring to mind affects their likelihood of leading to SI calculation. This study indeed found that the more likely people are to ask a polar question involving the stronger scalar (e.g., *Is the movie excellent?*), the higher the rate of SI calculation, but with the caveat that this correlation only holds for bounded scales.

Some properties of scales were hypothesized to predict scalar diversity, but were found not to. Sun et al. (2018) investigated the factor of scale homogeneity, testing the prediction that whenever a scalar term if polysemous, SI rates will be lower. To take an example, *hard* can be understood as belonging to the <*hard, unsolvable*> scale (which is what van Tiel et al. 2016 tested), but it is also possible to understand it as related to other scales, e.g. <*hard, unbearable*>. Less homogeneous scales are thus predicted to lead to lower SI rates. This factor was operationalized via a naturalness rating experiment of sentences such as *The student is brilliant but not intelligent*, where higher naturalness indicated lower homogeneity, predicting lower SI rates. However, the authors found that this factor only predicted scalar diversity insofar as it related to semantic distance.

In addition to distinctness, van Tiel et al. (2016) also put forth a hypothesis about the availability of the stronger scalar term. Availability is relevant because, in order for SI to arise, hearers must assume that the speaker considered using a stronger alternative (e.g., *all*) to what she ultimately uttered (*some*). As measures of availability, the authors considered: association strength between the weaker and stronger terms (measured via a modified cloze task), grammatical class, frequency (of the stronger term itself and the stronger term relative to the weaker term), and semantic relatedness (derived from distributional semantics, specifically Latent Semantic Analysis, Landauer and Dumais 1997). However, none of these was found to be a significant predictor of SI rates in the study. In later work, however, Westera and Boleda (2020) showed that a sufficiently fine-grained notion of semantic similarity (or relatedness) does actually affect SI rates. To measure similarity, these authors used the ELMo neural network model, which uses a different context window during training than Latent Semantic Analysis (sentence-length rather than document-length). The findings of this study also suggest that a context-dependent measure of semantic similarity can improve model performance, but only modestly. Interestingly, these authors found a negative correlation between semantic similarity and SI rates, not a positive correlation as van Tiel et al. (2016) predicted. In other words, Westera and Boleda (2020) found that the more semantically similar two scalar terms are, the lower the SI rate. They argue that this is because semantic relatedness in fact indexes distinctness: the more similar two terms are, the less distinct they are, and hence the lower the likelihood of SI.

Importantly for this dissertation, 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. While Westera and Boleda's (2020) results did reveal an effect of semantic relatedness (contra van Tiel et al. 2016), this metric still only captured 4-6% of the variance. Lastly, Pankratz and van Tiel (2021) found that relevance explained 4%. Models that include multiple known predictors from different studies still fall short of explaining all of the variance in SI rates: Sun et al. (2018) report that their best fitted model explained 63% of the variance, Gotzner et al.'s (2018) model explained 66%, while Pankratz and van Tiel (2021) report that their model combining relevance with other predictors explained 8%. In other words, a lot of scalar diversity is still unexplained.

4.2 New set of scales

The main motivation behind studies of scalar diversity is that claims regarding the calculation, processing, or any other properties of SI should not be made on the basis of a very small number of scales; instead, we should probe what properties can generalize to all possible lexical scales. Yet, in previous work on scalar diversity, the set of scales studied skewed towards, or concentrated entirely on, adjectives: e.g. van Tiel et al.'s (2016) set of scales contained 70% adjectives, and Gotzner et al.'s (2018)'s consisted entirely of adjectives. If our goal is to identify properties of SI that hold generally, across all scales, then it stands to reason that our empirical domain should not be (largely) restricted to a specific grammatical class, and we should instead investigate other open class scales e.g. verbal or adverbial scales as well. A smaller, additional worry is that it can be brought into question whether some of the items studied in previous work were in fact lexical scales at all. For example, on the putative (<*cheap, free*>) scale, *free* in fact entails not having a price, while *cheap* entails having a price, suggesting that these two scalar terms do not form a scale —which could then explain why van Tiel et al. (2016) found extremely high rates of endorsement of the conclusion that what is meant by *cheap* is *not free*.

To address the above problems, I constructed a new scale set, consisting of 60 lexical scales,

Adjective	*<allowed, obligatory>*; *<attractive, stunning>*; *<big, enormous>*; *<cool, cold>*; *<dark, black>*; *<difficult, impossible>*; *<dirty, filthy>*; *<funny, hilarious>*; *<good, excellent>*; *<happy, ecstatic>*; *<hard, unsolvable>*; *<harmful, deadly>*; *<hungry, starving>*; *<intelligent, brilliant>*; *<intimidating, terrifying>*; *<old, ancient>*; *<overweight, obese>*; *<palatable, delicious>*; *<polished, impeccable>*; *<possible, certain>*; *<pretty, beautiful>*; *<scared, petrified>*; *<serious, life-threatening>*; *<similar, identical>*; *<small, tiny>*; *<snug, tight>*; *<tired, exhausted>*; *<ugly, hideous>*; *<understandable, articulate>*; *<unpleasant, disgusting>*; *<warm, hot>*; *<willing, eager>*
Verb	*<begin, complete>*; *<believe, know>*; *<damage, destroy>*; *<dislike, loathe>*; *<double, triple>*; *<like, love>*; *<match, exceed>*; *<permit, require>*; *<reduce, eliminate>*; *<slow, stop>*; *<start, finish>*; *<survive, thrive>*; *<tolerate, encourage>*; *<try, succeed>*; *<want, need>*
Adverb	*<equally, more>*; *<here, everywhere>*; *<largely, totally>*; *<mostly, entirely>*; *<once, twice>*; *<overwhelmingly, unanimously>*; *<partially, completely>*; *<primarily, exclusively>*; *<probably, necessarily>*; *<usually, always>*; *<well, superbly>*
Connective	*<or, and>*
Quantifier	*<some, all>*

Table 4.2: 60 lexical scales collected.

which will form the basis of scalar diversity investigations in this dissertation. The method for compiling this set was as follows. Scales used in van Tiel et al. (2016)'s study, as well as those used in de Marneffe and Tonhauser (2019)'s were taken, and this was supplemented by corpus data. In particular, I conducted corpus searches in the Corpus of Contemporary American English (COCA, Davies 2008), searching for the following: *X or even Y*; *not just X but Y*; *X but not Y*. These searches were conducted for adjectives, verbs, and adverbs. The expectation is that these searches would largely uncover sentences from the corpus where a lexical scale was produced; in particular, scales where X is the weaker scalar term and Y is the stronger scalar term. Sentences where X and Y were clearly not in a scale-mate relation were discarded —this was done based on researcher intuition, eliminating corpus results such as *unreasonable or even bloodthirsty*. Combining the items from the two published studies with the corpus data resulted in a total number of 101 items. As the next step, the following semantic tests were conducted to probe whether X and Y indeed form a scale:

- Is *X and even Y* odd? —Expected answer: No

- Is *X but not Y* contradictory? —Expected answer: No
- Is *Y but not X* contradictory? —Expected answer: Yes

The *and even* test is for cancellability: if the *not Y* inference arising from *X* is an SI, it should be cancellable, that is, *Y* should be assertable (Grice, 1967). The *but not* tests probe for asymmetric entailment (Horn, 1972): Y should entail X, but not vice versa, for X and Y to qualify as scale-mates. Wherever a pair did not produce the expected answer, it was excluded. Lastly, wherever one word participated in more than one scale, one of those scales was excluded, e.g. because *good* occurred in both the <*adequate, good*> scale and the <*good, excellent*> scale, the former scale was excluded. The reason for this was that in lexical priming experiments (Chapter 2), participants should not be exposed to critical experimental lexical items more than once.

Overall, the scale collection procedure resulted in a final set of 60 <*weaker, stronger*> scalar terms, which can be seen in Table 4.2.

4.3 Experiment 7: Replicating scalar diversity

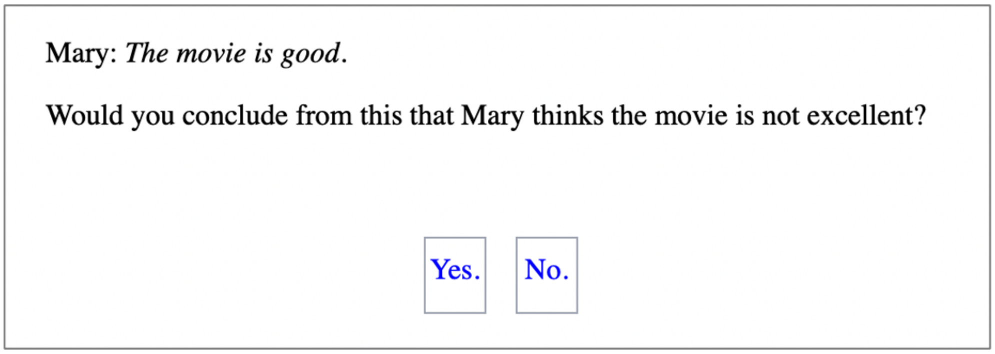
In the following, I report on an experiment testing the rates of inference calculation from the 60 scales discussed above.

Participants

42 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond, 2007). 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. 1 participant was removed from analysis because the background questionnaire revealed that they were bilingual, and 1 additional participant was removed based on having a reaction time shorter than 500ms on the majority of the trials, as well as answering "No" in the first half and "Yes" in the second half of the trials, suggesting that they were not paying attention to the task. Data from 40 participants is reported below.

Task, materials and procedure

An inference task was used to investigate the likelihood of deriving an SI from 60 different scales. Participants were presented with a sentence such as “Mary: *The movie is good.*” and were asked the question “Would you conclude from this that Mary thinks the movie is not excellent?”. They responded by clicking “Yes” or “No”. Figure 4.1 shows an example trial item.



Mary: *The movie is good*.

Would you conclude from this that Mary thinks the movie is not excellent?

Figure 4.1: Example experimental trial from Experiment 7, which measured SI rates in a verification task.

A “Yes” answer indicates that the participant has calculated the relevant SI (*good* → *not excellent*), while a “No” answer indicates that the participant has not calculated the SI, i.e. they are interpreting *good* as meaning *at least good*.

7 filler items were also included, which contained two terms that are either in an entailment relation (*wide* → *not narrow*), or unrelated (*sleepy* → *not rich*). Given that the filler items had a clear, correct “Yes” or “No” answer, they 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 67 trials.

Predictions

Given consistent findings of scalar diversity in existing literature, I predict robust variation across the 60 different scales in how likely they are to lead to SI calculation. That is, I predict that the percentage of “Yes” vs. “No” responses in the inference task of Experiment 7 will vary substantially from scale to scale.

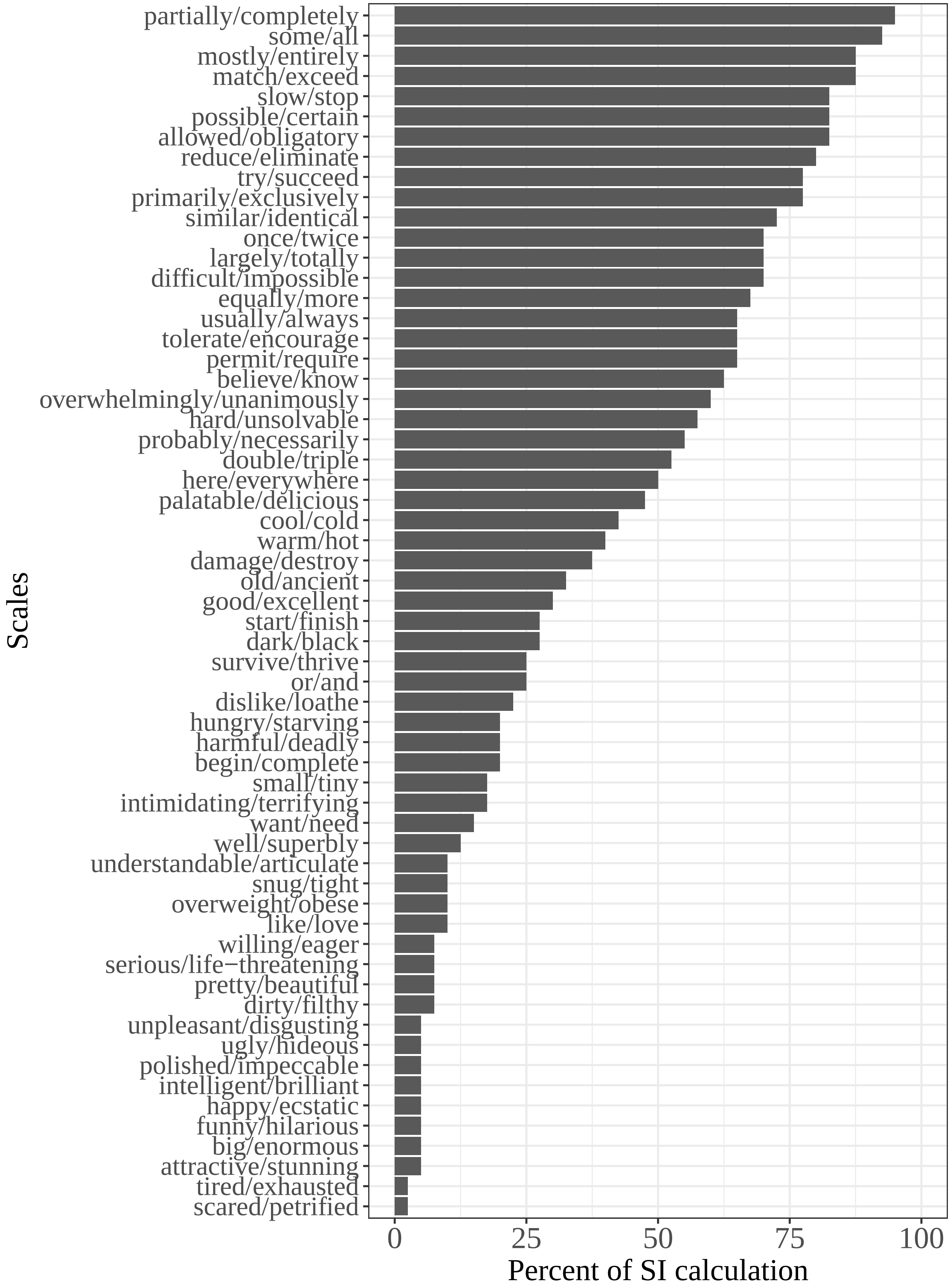


Figure 4.2: Results of Experiment 7: Inference rates for 60 different scales.

Results

Figure 4.2 shows the results of Experiment 7. Percent of inference calculation corresponds to the proportion of "Yes" responses. As is evident from this figure, considerable variation was found among critical items. In particular, positive responses, i.e. the rate of SI calculation, range along a continuum from 2.5% (for <*scared, petrified*> and <*tired, exhausted*>) to 95% (for <*partially, completely*>). This result thus successfully replicates the scalar diversity effect: different scalar expressions yielded wildly different rates of SI.

4.3.1 Quantifying scalar diversity

While the observation of scalar diversity was based only on descriptive statistics in previous work (range of SI rates), in this chapter I propose a more rigorous measure to quantify the variation across scales. Specifically, I turn to information theoretic measures, which are commonly used, for instance, in the domain of syntactic processing (e.g., surprisal: Levy 2008; entropy reduction: Hale 2003). In particular, I propose using relative entropy (Kullback and Leibler, 1951), a measure that compares two probability distributions and quantifies their difference. To quantify scalar diversity in Experiment 7, I treated the normalized SI rates (i.e., percentage of "Yes" responses) across different scales as a probability distribution. I then compared this distribution to the uniform distribution. The uniform distribution represents a (hypothetical) scenario where each scale leads to the same SI rate. This scenario reflects the implicit assumption made by theoretical accounts, which suggest that SI calculation, which proceeds via reasoning about a stronger scalar alternative, should not vary across different scales—the so-called "uniformity assumption" (van Tiel et al., 2016, p. 139). As I quantify diversity via comparison to a uniform distribution, I do not assume any particular SI rate as the basis for uniformity; in my calculation of relative entropy, I remain agnostic about whether "uniform" would mean 100% SI rate across all scales, 70%, or so on. Intuitively, relative entropy represents how "surprised" we are if we assume a particular distribution (the uniform distribution), but observe a different one (Experiment 7).

The equation in (7) was used to calculate relative entropy. Here, $p(x)$ is the normalized observed percentage of "Yes" responses across scales in Experiment 7, $\mathscr{X}$ is the 60 scales, i.e., the finite set over which I defined the probability distribution, and $q(x) = 1/60$ is the uniform probability mass function over the 60 scales. In this specific case, because the uniform inference rate is a constant across all 60 scales, the relative entropy that we obtain is the entropy of the uniform distribution minus the entropy of the experimentally collected SI rates.

(7) Let $p(x)$ and $q(x)$ be probability mass functions over the same set $\mathscr{X}$. The relative entropy of $p(x)$ with respect to $q(x)$ is given by

$$D(p||q) = \sum_{x \in \mathscr{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

The relative entropy of Experiment 7 SI rates is 0.466. To contextualize this number, we may consider a number of hypothetical scenarios as benchmarks. If all scales indeed led to the same rate of SI calculation, then that would give a relative entropy of 0—see the Benchmark 1 facet in Figure 4.3. At the other extreme, the highest possible relative entropy would be obtained if all the probability mass was concentrated on a single scale: that is, if only one of the 60 scales ever led to SI calculation (at some non-zero rate), while the other 59 scales did not—this hypothetical scenario would lead to a relative entropy of 5.907, and it is shown as Benchmark 2 in Figure 4.3. Closer to the actual experimental findings is Benchmark 3: a hypothetical "linear" distribution, where likelihood of SI calculation is evenly distributed across the 60 lexical scales over a 0-100 range. Here, for instance, one scale leads to SI calculation at a 1.67% rate, the next at 3.33%, the one after that at 5%, etc., up to scale number 60 leading to SI calculation at 100%. This linear benchmark would yield a relative entropy of 0.2912. Lastly, the "quadratic" distribution in Benchmark 4 is a scenario similar to Benchmark 3, in that every scale has a unique SI calculation rate; but here, probability mass is more concentrated toward one scale, giving a relative entropy of 0.6352. We can see that the experimentally collected rates fall between Benchmarks 3 and 4 (0.466), suggesting more diversity than Benchmark 3, but less than 4. Table 4.3 summarizes these results.

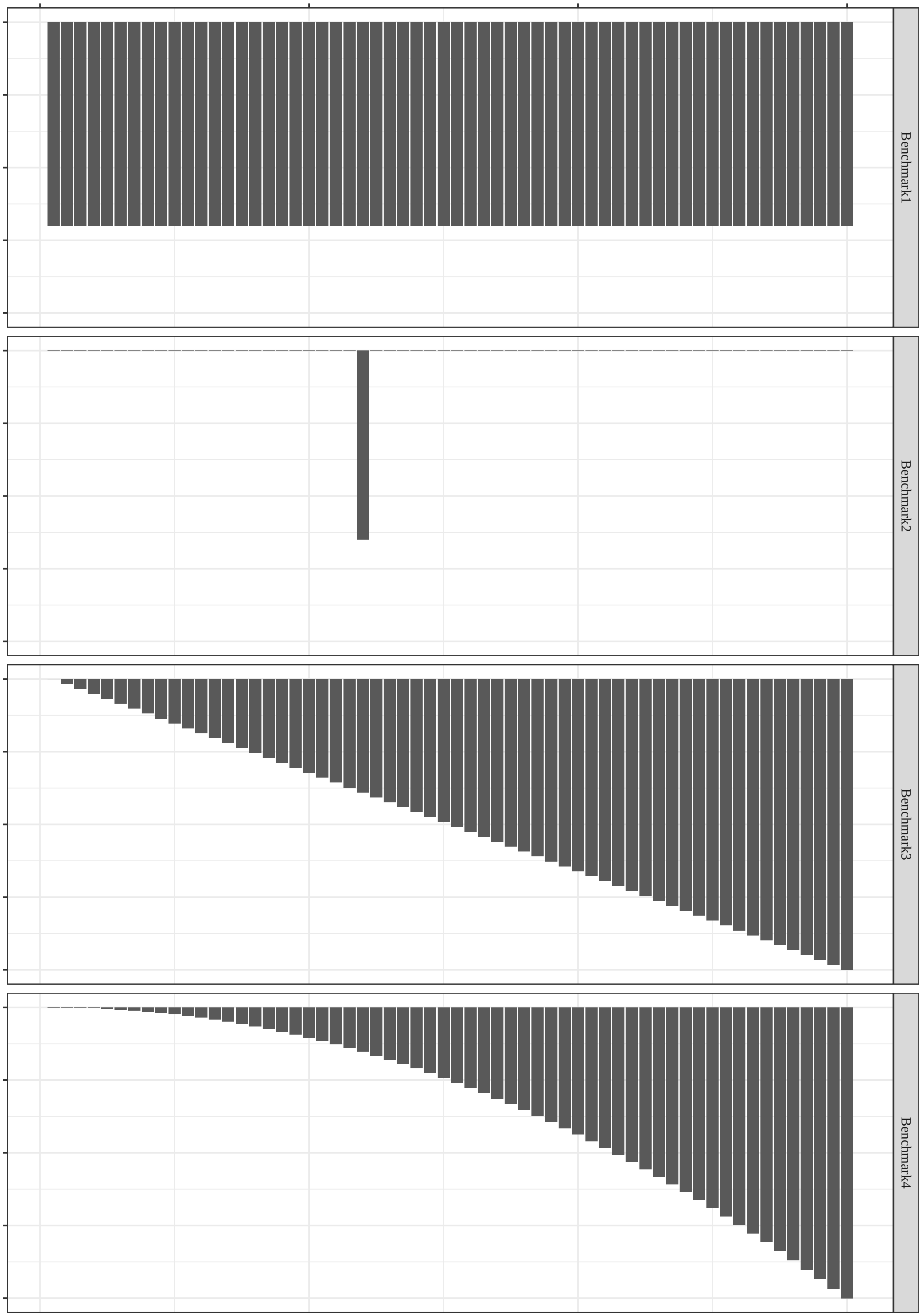


Figure 4.3: Hypothetical SI rates: benchmarks for quantifying diversity using relative entropy.

Condition	Relative entropy
Experiment 7	0.466
Benchmark 1	0
Benchmark 2	5.907
Benchmark 3	0.2912
Benchmark 4	0.6352

Table 4.3: Relative entropy results from Experiment 7 and the four hypothetical benchmarks

The benchmarks outlined here are for illustration; my main goal will be to use the proposed relative entropy measure to compare different sets of experimentally collected SI rates to one another, seeing how different manipulations reduce variation across lexical scales. This is what we will explore in Chapter 5.

4.4 Explaining scalar diversity

In the following, I detail the motivation for, and operationalization of, different factors that may be predicted to capture the observed variation in SI rates across scales. Specifically, we will look at:

- the accessibility of the stronger alternative (Section 4.4.1),
- the distinctness of the weaker and stronger scalar term (Section 4.4.2),
- boundedness of the scale (Section 4.4.3),
- (relative/absolute) frequency of the stronger alternative (Section 4.4.4),
- semantic similarity of the weaker and stronger scalar term (Section 4.4.5),
- the meaning of the negated stronger alternative (Section 4.4.6).

4.4.1 Experiment 8: Accessibility of stronger alternative

In Experiment 8, I used a modified cloze task as a metric for the accessibility of stronger alternatives across scales, which was found to significantly predict the rates of SI calculation from Experiment 7.

Hypothesis

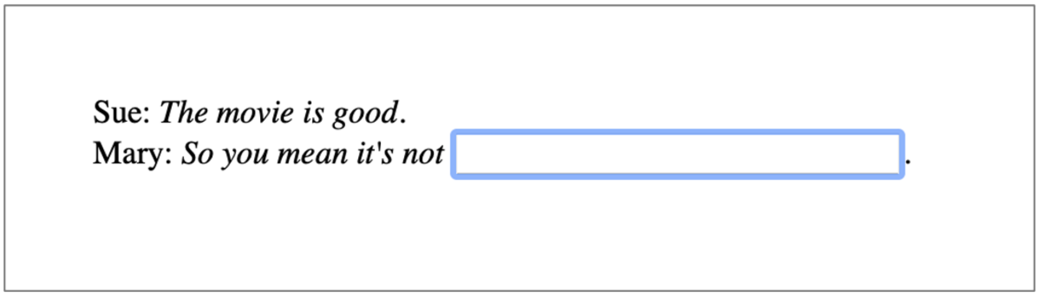
I hypothesize that scalar diversity can, in part, be explained by how accessible a stronger alternative is, given the weaker scalar. The causal mechanism behind this hypothesis is as follows. I assume that SI calculation proceeds via reasoning about alternatives, and that hearers generate a set of alternatives when they encounter a potentially SI-triggering utterance that contains a weak scalar term. The more accessible an alternative is, the more likely hearers are to reason about it, and therefore the more likely the relevant SI is to arise. In the context of scalar diversity, the intuition is that there may be differences across scales in how strongly the weaker scalar evokes a stronger alternative. For instance, it is possible that when encountering a sentence containing *some*, the stronger alternative *all* always comes to mind; but when encountering a sentence containing *good*, a number of competing alternatives may be activated, such as *excellent, funny, thrilling, thought-provoking*, and so on.

Participants, task and procedure

61 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 7. Data from all 61 participants is reported below.

I operationalize alternative accessibility as cloze probability, a commonly used measure of the predictions the parser makes in language comprehension. In particular, the probability of a target word completing a given sentence frame is taken to index how expected a word is in a context (Taylor 1953; see also i.a. Kutas and Hillyard 1984). My experiment employed a modified cloze task: participants were presented with a dialogue context where Sue uttered a potentially SI-triggering sentence, such as *The movie is good* (identical Experiment 7), and Mary followed up by saying *So you mean it's not BLANK*. Participants were instructed to complete Mary's utterance with the first word that came to mind, making sure that their completion made sense in the context of the dialogue. Figure 4.4 shows an example trial item.

Similarly to Experiment 7, Experiment 8 included 60 experimental trials and 7 fillers to serve



Sue: *The movie is good.*
Mary: *So you mean it's not* ______.

Figure 4.4: Example experimental trial from Experiment 8

as catch trials. The 2 practice trials at the beginning of the experiment also provided participants with some feedback on what is a reasonable completion in the cloze task. Experiment 8 included two within-participants conditions that addressed different research questions and are not discussed here; due to counterbalancing, I therefore ended up collecting 19-22 completions per scale. The experiment was administered in a Latin Square design.

Prediction

Given the accessibility hypothesis, the prediction I make for the results of Experiment 8 is that the more frequently the stronger alternative (e.g., *all*) is mentioned in the modified cloze task, the higher the SI rate for that scale (*some but not all*) from Experiment 7.

Results and discussion

The results of the cloze task were coded as follows. I counted the number of times the relevant alternative was mentioned, and divided this by the number of total completions for that scale, resulting in the percent of stronger alternative mentioned. Figure 4.5 shows these results, correlated with SI rates from Experiment 7. In the coding of the results, synonyms of the stronger alternative were also counted. There was a positive correlation between the results of Experiment 8 and 7 (Pearson's correlation test: r=0.59, p<0.001). The higher the percent of mentioning the stronger alternative (*excellent*) in the cloze task, the higher the corresponding SI rate from that scale (*good but not excellent*). For a more detailed statistical analysis, combining all experiments and analyzing

SI calculation as “Yes” vs. “No” responses, see Section 4.4.7.

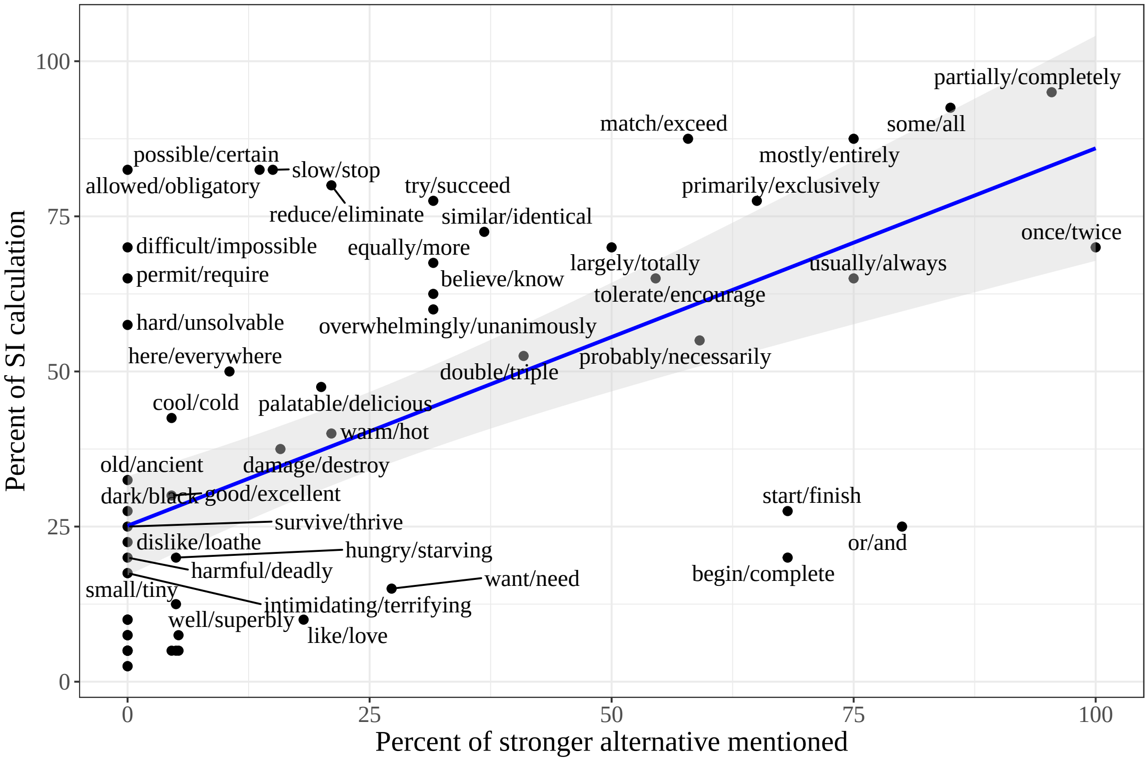


Figure 4.5: Results of Experiments 7 and 8: The *x* axis shows alternative accessibility from Experiment 8. The *y* axis shows SI rate from Experiment 7.

In other words, scalar diversity was shown to be predicted by the accessibility of the stronger scalar —that is, by how strongly a weaker scalar evokes a stronger alternative. To provide a few illustrative examples beyond the overall quantitative analysis, for some scales in Experiment 8, the stronger alternative was given almost every time as a cloze completion; for instance, for *some*, the alternative *all* was provided by almost all participants (with one participant providing *most*). On the opposite end, for some scales the stronger alternative from Experiment 7’s inference task was never provided: for instance, for *good*, the completions included *bad, terrible, overrated*, but not *excellent*. This suggests that the relevant stronger alternative is not very accessible here. Impressionistically, in such cases, antonyms to the weaker scalar term were frequent completions. Lastly, some scales led to a greater variety of completions, suggesting that a larger number of (not just scalar) alternatives can be activated upon encountering the SI-triggering utterance: for *try*, for example, participants filled in the stronger alternative *succeed*, but also *fail, surrender, concede, quit*.

A potential caveat to mention is that my measure of accessibility may be interpreted as the production side of scalar diversity. In the inference task of Experiment 7, participants have to judge statements containing the negated stronger scalar (*not excellent*), having seen an SI-triggering sentence. In Experiment 8, participants were asked to fill in a blank under negation (*So you mean it's not...*), as a response to the same SI-triggering sentences. Perhaps the reason that the results of the cloze task predict SI rates, and both experiments show diversity, is that we are tapping into outcomes of the same mechanism, with Experiment 7 testing the comprehension and Experiment 8 the production side. However, there is one important difference I would like to highlight: the inference task asks participants to make a decision about a particular stronger alternative. The cloze task, on the other hand, probes what is a relevant contrast for a weaker scalar term—e.g., is it *good* vs. *excellent*, or *good* vs. *bad*. Ultimately, the cloze task is therefore informative regarding whether the specific alternative message that the hearer infers the negation of – having seen an SI-triggering utterance – is the same as the stronger alternative from the lexical scale that we test in the inference task.

My proposal of alternative accessibility is closely related to van Tiel et al.'s (2016) proposal that the availability of the stronger alternative should predict scalar diversity. The authors argue that for SI to arise, it has to be the case that the speaker could have actually considered using the stronger scalar term instead of the weaker one she uttered. As mentioned in Section 4.1.2, van Tiel et al. (2016) tested four different operationalizations of availability, but none of them were found to be a predictor of diversity. My operationalization is novel in that it utilizes a language production task in a discourse context, and it does end up predicting SI rates.

4.4.2 Experiment 9: Distinctness of scalar terms

In Experiment 9, I used posterior degree estimates as a metric for the distinctness of the weak and strong scalar terms, which significantly predicted the likelihood of SI calculation.

Hypothesis

Distinctness of scalar terms was originally put forth by van Tiel et al. (2016) as a potential explanation for scalar diversity. Distinctness is relevant for the likelihood of SI calculation for the following reason. The inferential process underlying SI calculation involves the hearer reasoning about, and negating, a stronger alternative (*all*) that the speaker could have said, but did not. For this reasoning to go through, there has to be a clear stronger alternative, and it has to be sufficiently stronger. In other words, the more distinct two scalar terms (*some* vs. *all*) are, the more likely the hearer is to assume that the speaker should have used the stronger term if possible. If it is difficult to distinguish the weak and strong scalar, e.g. if they are near-synonyms, SI calculation is unlikely.

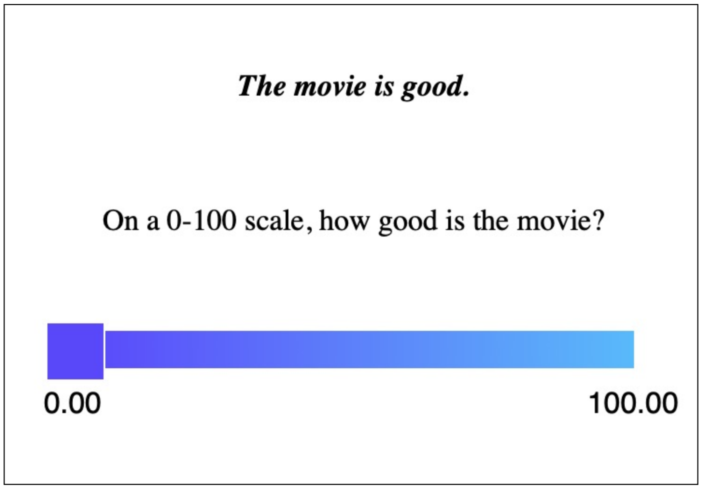
Participants, task and procedure

60 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 7. Data from all 60 participants is reported below.

To operationalize the distinctness of scalar terms, I experimentally collected degree estimates on the underlying scales. Specifically, in Experiment 9, I am interested in what world states hearers think utterances such as *The movie is good* vs. *The movie is excellent* describe. In other words, what this experiment tests is: after encountering the relevant utterance, what degree of goodness do hearers ascribe to the movie?

Experiment 9 therefore employed a degree estimate task. Participants were presented with a sentence such as *The movie is good* or *The movie is excellent*, and were instructed to answer a question like *On a 0-100 scale, how good is the movie?* by picking a point on a scale from 0 to 100. The weak and strong scalar terms were tested as a between-participants manipulations (30 participants in each condition). Figure 4.6 shows an example trial item from the weak scalar term condition and Figure 4.7 from the strong scalar term condition.

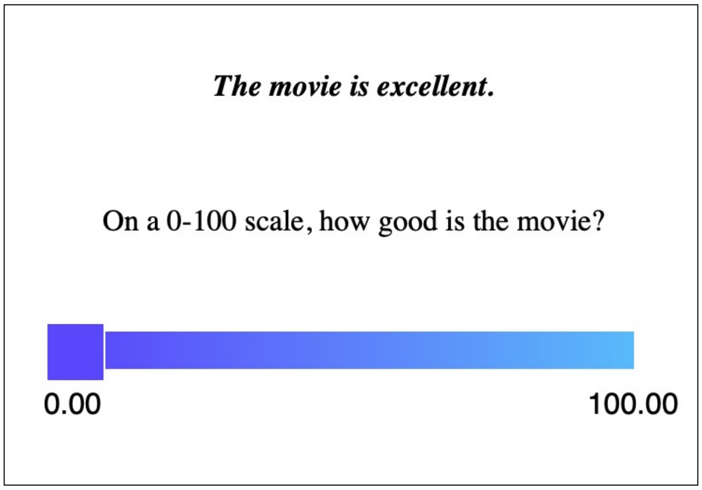
I aimed to create neutral questions that would not bias participants toward either end of the scale. For adjectival lexical scales, questions relied on the weaker term wherever possible (*On a*



The movie is good.

On a 0-100 scale, how good is the movie?

Figure 4.6: Example experimental trial from Experiment 9: weaker scalar term



The movie is excellent.

On a 0-100 scale, how good is the movie?

Figure 4.7: Example experimental trial from Experiment 9: stronger scalar term

0-100 scale, how old is the house? for <*old, ancient*>), while in other cases I picked a neutral underlying adjective, e.g., *On a 0-100 scale, how likely is success?* for <*possible, certain*>. Questions for verbal and adverbial scales were necessarily more varied, but aimed to be neutral and refer to the underlying scale, e.g., *On a 0-100 scale, how much will the sales increase?* for <*double, triple*> or *On a 0-100 scale, how often is the lawyer early?* for <*usually, always*>. This task is an idealization, because not all lexical scales map onto a bounded underlying degree scale, but results suggest that participants were able to accommodate and make sense of the task in the context of this experiment.

The experiment included 60 critical items, as well as 3 practice trials and 5 filler items. The latter served as catch trials and used antonyms in the sentence and task question, e.g., *The table is clean* was paired with *On a 0-100 scale, how dirty is the table?*.

Prediction

The data collected in Experiment 9 represents hearers' probabilistic guesses on what world state the speaker has in mind, given her utterance. Based on the distinctness hypothesis, I predict that the greater the difference between the degree estimates for the weak and the strong scalar terms, i.e. the further apart they are on the underlying degree scale, the higher the SI rate will be for that scale. As mentioned, this is because for an SI (*good but not excellent*) to arise, *good* and *excellent* have to be perceived as describing two different world states.

Results and discussion

Averaged over the 60 lexical scales, the stronger scalar terms received higher ratings than the weaker terms —see Figure 4.8. In other words, a sentence such as *The movie is excellent* led hearers to attribute a higher degree of goodness to the movie than *The movie is good*. This difference is statistically significant: using the lme4 package in R (Bates et al., 2015), I fit a linear mixed effects regression model that predicted Response (0-100) by Condition ("weak" and "strong"). The fixed effects predictor Condition was treatment-coded, with weak as the reference level. Random intercepts were included for participants and items. Responses to strong terms were found to be significantly higher than to weak terms (Estimate=22.68, Std. Error: 2.68, t=8.38, p<0.001). This serves as confirmation that participants were performing the task adequately.

To check the prediction of distinctness, I took the absolute difference in means between the weak and strong terms: for instance, *The movie is good* received a response of 69.4 on the 0-100 scale, while *The movie is excellent* received 89.1, resulting in a "distinctness" value of 19.7. Figure 4.9 shows these results, correlated with SI rates from Experiment 7. As can be seen in the figure, there was a positive correlation between the results of Experiment 9 and 7 (Pearson's correlation

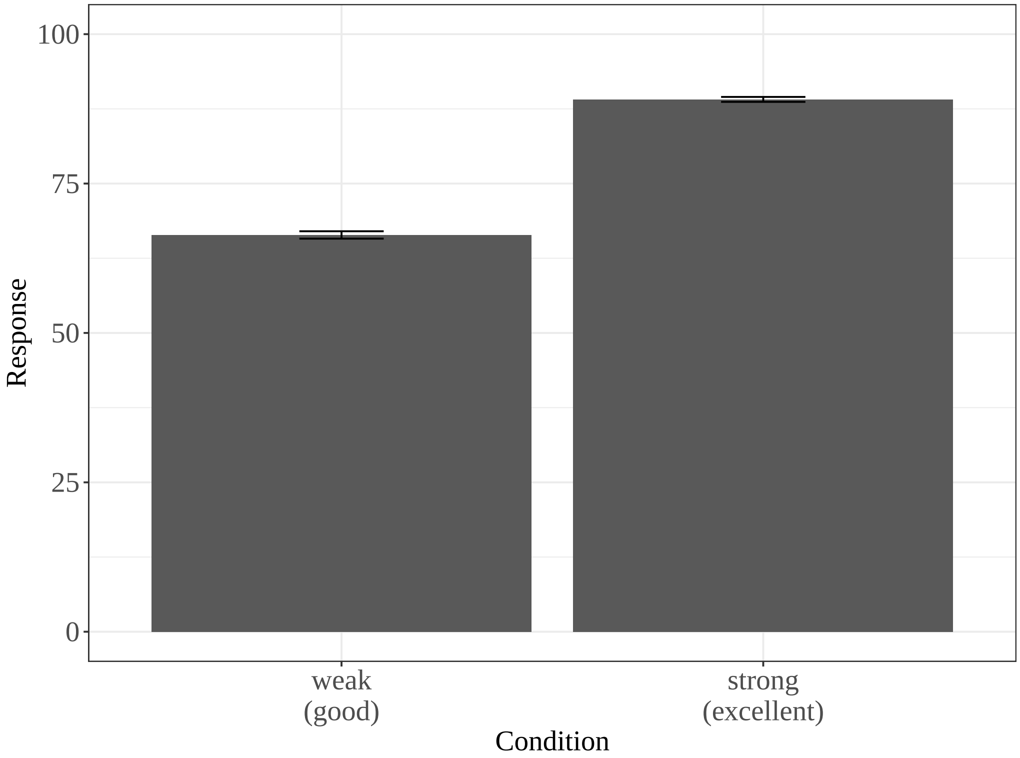


Figure 4.8: Results of Experiment 9: Average response on the degree estimate task (0-100 scale) for the two different conditions

test: r=0.33, $p<0.05$). That is, scalar diversity was shown to be predicted by the distinctness of scalemates. Specifically, the higher the (absolute) difference between a weak (*good*) and a strong (*excellent*) term, as measured via degree estimates, the higher the corresponding SI rate from that scale (*good but not excellent*).

In other words, I found that the more distinct the world states that the weaker and the stronger term are taken to describe, the higher the SI rate for that scale. Experiment 9's results thus present further evidence for van Tiel et al.'s (2016) distinctness hypothesis, using a novel operationalization that relies on empirically collected posterior degree estimates. Van Tiel et al. relied on the notion of boundedness, as well as experimentally collected judgements about semantic distance, to test the distinctness hypothesis. It is worth discussing how the latter relates to my Experiment 9. In the semantic distance experiment, participants were presented with a pair of sentences, such as *She is intelligent* and *She is brilliant*. They then had to respond to the question *Is statement 2 stronger than statement 1?* via a 7-point Likert scale, where 1 corresponded to "equally strong" and 7 to

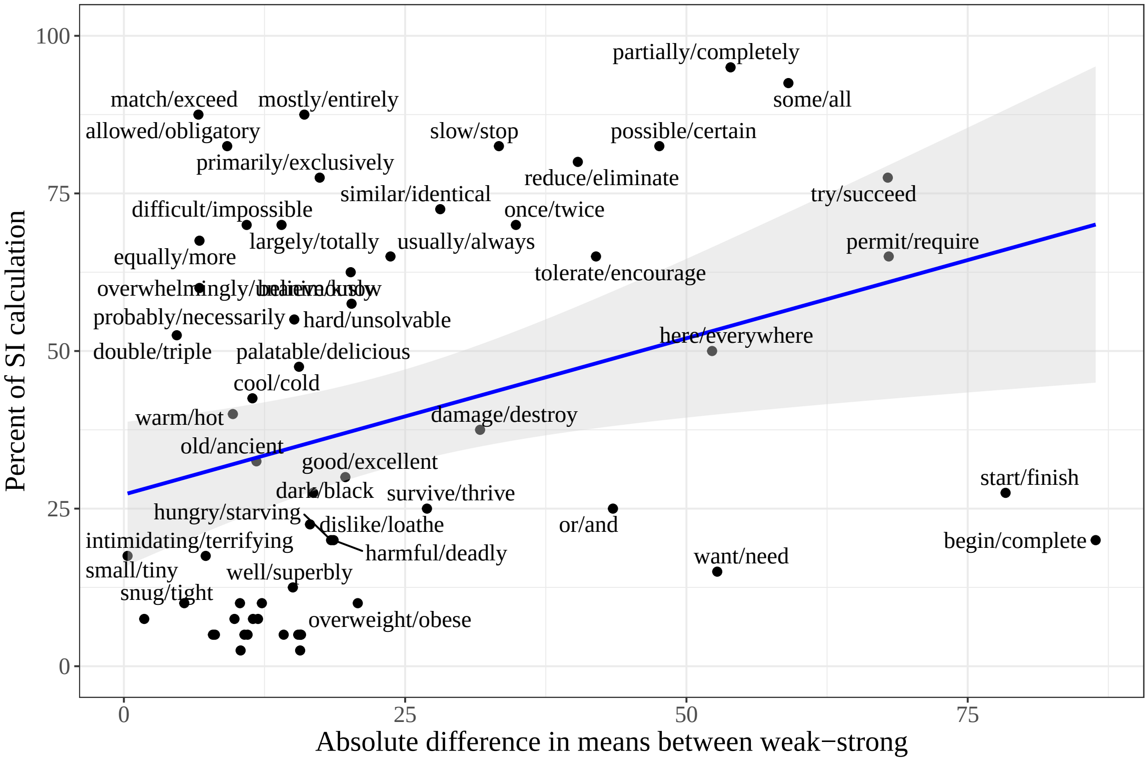


Figure 4.9: Results of Experiments 7 and 9: The *x* axis shows distinctness between each weak-strong scalar pair from Experiment 9. The *y* axis shows SI rate from Experiment 7.

"much stronger". In line with the distinctness hypothesis, the authors found that semantic distance was positively correlated with SI rates: the more distant a weak and a strong scalar term were in their experiment, the more likely the corresponding SI. My Experiment 9 differs from van Tiel et al.'s in that it does not assume any *a priori* strength relation; my experimental instructions did not presuppose that one statement could be stronger than the other, and participants simply picked points on an underlying scale. Another notable difference is that judging the relative strength of statements requires a metalinguistic judgment, and therefore degree estimates are arguably a more natural task. Altogether, the experiment reported here constitutes further evidence for van Tiel et al.'s distinctness hypothesis, going beyond existing evidence in the prior literature.

4.4.3 Boundedness

Van Tiel et al. (2016) define a scale as bounded if the stronger scalar denotes an endpoint (see also Kennedy and McNally 2005); <*some, all*> is therefore an example of a bounded scale. On the other hand, in unbounded scales, both scalar terms denote intervals; <*good, excellent*> is is an unbounded scale. Thus we can see that boundedness depends only on the semantics of the stronger scalar term. The 60 scales investigated in this dissertation were categorized as bounded or unbounded given these definitions: 26 scales were bounded and 34 unbounded.

For van Tiel et al. (2016), boundedness is a component of distinctness; they hypothesize that that scalar terms on bounded scales are more distinct: in other words, they are easier to distinguish than pairs of scalar terms on non-bounded scales. This is because the scalar terms on bounded scales can be distinguished on formal grounds, since one denotes an endpoint and the other denotes an interval. In the case of unbounded scales, however, to distinguish the two scalar terms, one needs to inspect the exact domain their respective intervals cover.

An alternative, but related way of conceptualizing the relevance of boundedness for SI calculation is as follows. In bounded scales, the stronger scalar term denotes an endpoint, which makes it very salient as an alternative to the vague, weaker term. This high level of salience for the stronger alternative is what leads to higher rates of SI calculation, as compared to unbounded scales, where both terms are vague (Ronai and Xiang, 2021). Nonetheless, both potential accounts spelled out here make the same prediction, that bounded scales should lead to more SI calculation.

Figure 4.10 shows SI rate results split according to boundedness. For the statistical analysis, a logistic mixed effects regression model (lme4 package in R, Bates et al. 2015) was fit to compare the rates of SI calculation between bounded and unbounded scales. The model predicted Response ("Yes" vs. "No" in the Experiment 7 inference task) as a function of Boundedness. It included the maximal random effects structure supported by the data (Barr et al., 2013): random slopes and intercepts for participants and random intercepts for items. The fixed effects predictor Boundedness (bounded vs. unbounded) was sum-coded before analysis ("bounded": 0.5 and "unbounded": 0.5). This analysis revealed overall lower SI rates for unbounded, as compared to bounded scales

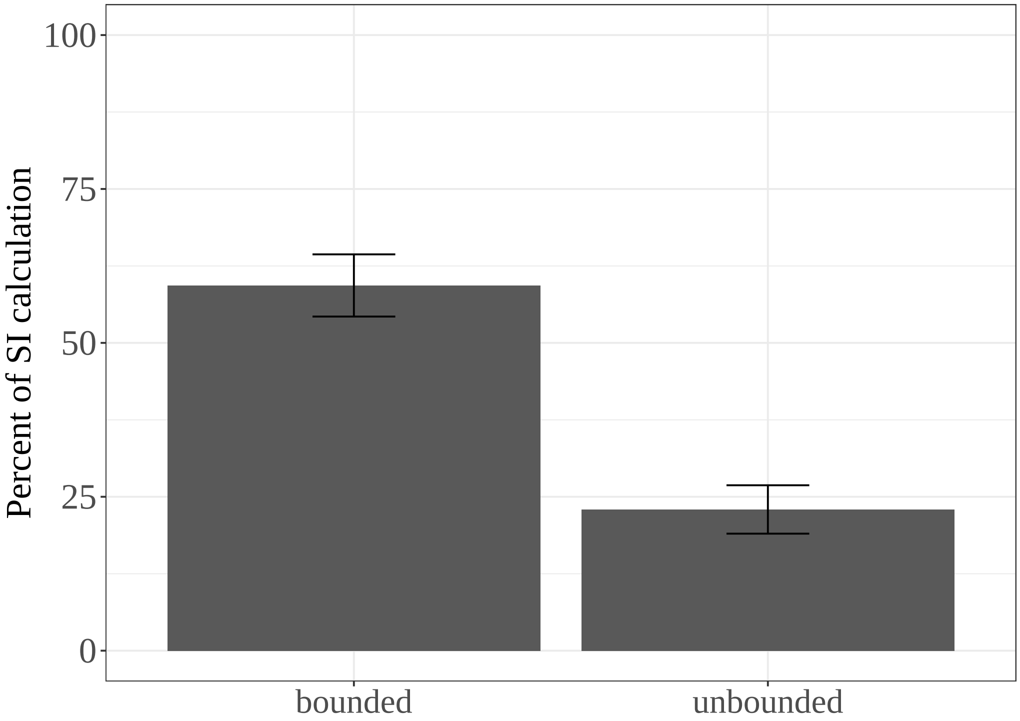


Figure 4.10: Results of the boundedness analysis: The *x* axis compares bounded and unbounded scales. The *y* axis shows SI rate from Experiment 7.

(Estimate=-2.23, SE=0.41, *z*=-5.45, *p* <0.001). I therefore successfully replicate van Tiel et al.'s boundedness finding on a new scale set.

4.4.4 Frequency

Van Tiel et al. (2016) hypothesized that the availability of the stronger scalar term might predict the observed variation in SI rates: the more available the stronger scalar, the higher the SI rate. One way in which availability was operationalized is frequency. Though van Tiel et al. (2016)'s study ultimately found frequency not to be a predictor of scalar diversity, given that this dissertation uses a (partially) different set of scales with more variation in grammatical category, it is worth revisiting this analysis.

The reasoning for why frequency should matter is as follows. SI arises if the hearer has reason to believe that a stronger scalar term S was available to the speaker, but they chose to utter the weaker scalar term W instead. If, in a context where a speaker used W, S is quite frequent, then it should be available to the speaker as an alternative. On the other hand, if S has a low frequency, then this can explain why the speaker did not use it, even if it would have been appropriate and led to a true statement. To calculate the relative frequency of the stronger scalar term, I extracted the frequencies of all 120 scalar terms (for the 60 scales used in Experiments 7) from COCA (Davies, 2008). For each scale, the frequency of the stronger scalar term was divided by the frequency of the weaker scalar term. In order to reduce how skewed the resulting distribution was, the outcome of this division was log-transformed. The prediction is that availability, and in turn SI rate, is an increasing function of the frequency of S (relative to that of W). This prediction was, however, not borne out: relative frequency did not significantly correlate with SI rate (Pearson's correlation test: r=-0.11, p =.42).

As van Tiel et al. (2016) highlights, another possibility is that it is not the relative frequency

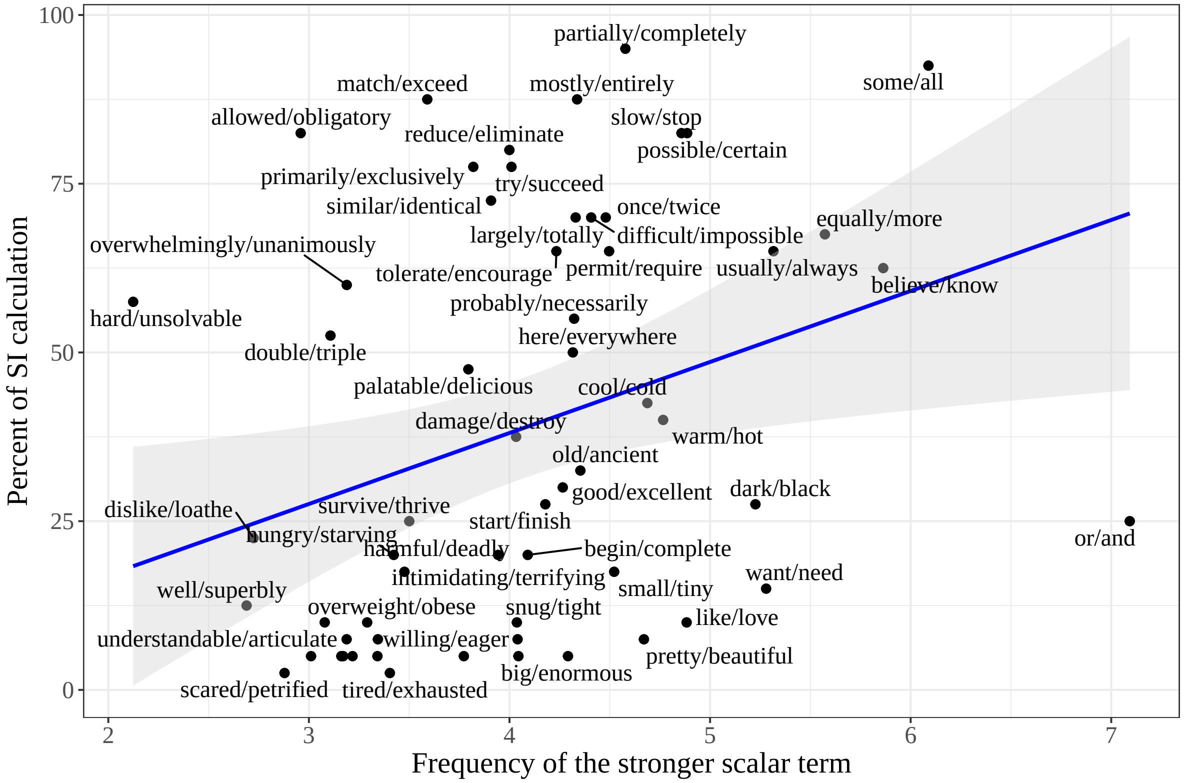


Figure 4.11: Results of the frequency analysis: The x axis shows the log-transformed frequency of the stronger scalar term. The y axis shows SI rate from Experiment 7.

of the stronger scalar term that matters, but rather its absolute frequency. In this case, the intuition is that the speaker's likelihood of considering S depends on how frequent a word S is, and this likelihood might be low if S is overall a very infrequent word, even if it might be more frequent than W. To test this hypothesis, the log-transformed frequencies of the stronger scalar terms were taken as the absolute frequency metric. I found a positive correlation between the frequency of the stronger scalar and SI rates (Pearson's correlation test: r=0.32, $p<0.05$) —see Figure 4.11. This finding thus provides support for van Tiel et al. (2016)'s hypothesis (though their own data did not). In further analyses, I therefore use the absolute frequency of the stronger scalar term as the index of frequency.

In van Tiel et al. (2016)'s results, neither relative, nor absolute frequency were found to correlate with SI rates. The scale set tested in this dissertation is therefore different in that absolute frequency does predict SI rates. However, as will be shown in Section 4.4.7, once we take other predictors into account, the effect observed for frequency goes away.

4.4.5 Semantic similarity

Semantic similarity, or semantic relatedness, was predicted by van Tiel et al. (2016) to positively correlate with SI rate: the more similar the weaker and stronger terms are to each other, the higher the SI rate. This was one way to operationalize the availability of the stronger scalar: for that scalar term to be available as an alternative the speaker could have uttered, the two scalar terms should be relevant, and therefore used, in similar contexts —which corresponds to them having similar semantic representations in distributional semantics (Harris, 1954).

While van Tiel et al. (2016)'s reasoning was that similarity correlates with availability such that the more similar the two scalar terms are, the higher the SI rate, the opposite prediction can also be made. Semantic similarity can also be taken to contribute to the distinctness of the two scalar terms: the more similar they are, the less distinct they are, meaning that the non-use of the stronger scalar can be attributed to it not being sufficiently different from the weaker term. This reasoning predicts semantic similarity to negatively correlate with SI rates: the more similar the scalar terms are, the less distinct they are, and therefore the lower the SI rate —which is in fact what Westera and Boleda (2020) found. Overall, it is worth revisiting the metric of semantic similarity even though van Tiel and Schaeken (2017) found it not to be a predictor of scalar diversity. This is because this dissertation is looking at a different set of scales, and also because more recent work indicates that some measures of semantic similarity are better predictors than the others (Westera and Boleda, 2020).

As already alluded to, semantic similarity will be quantified based on distributional semantics, or more specifically the "distributional hypothesis" (Harris, 1954). This hypothesis states that words that have similar meanings are used in similar linguistic environments—that is, they have similar distributions. Word meanings are represented by embeddings, encoded as a numerical vector. These are derived by abstracting over occurrences of the words in large sets of data. The semantic similarity of two words is computed as the cosine of the angle between the two vectors corresponding to the words —commonly called cosine similarity. The cosine is 1 when the vectors go in the same direction (high similarity), 0 when the vectors are orthogonal, and -1 when they go in the opposite direction (low similarity). For further details, the reader is referred to Chapter 6 (Section 6.4) of Jurafsky and Martin (2020). This dissertation uses three specific measures of semantic similarity, to be detailed below.

Latent semantic analysis

Latent semantic analysis (LSA) is a so-called "count-based" method of distributional semantics (Landauer and Dumais, 1997). Such a method starts from a large table of words from a corpus, which appear as both rows and columns. A row contains binary values (1 or 0), which represent whether the two words occur in the same sentence: if the two words both occur in a sentence, this is reflected by a 1 in the same column. Based on this matrix, dimensionality reduction is applied to calculate the word vectors. LSA values therefore reflect how often two words co-occur with the same words. For this analysis, I used the tool provided at http://lsa.colorado.edu/, in particular the 'General Reading up to 1st year college' as topic space (pairwise comparison), to extract LSA values for the 120 scalar terms.

LSA values were positively correlated with SI rates (Pearson's correlation test: r=0.26, p<.05) —see Figure 4.12. This finding is in line with van Tiel et al.'s (2016) hypothesis that the more semantically similar two scalar terms are, the more available the stronger one is as a potential alternative, and consequently the higher the SI rate —though ultimately the authors' own study did not provide evidence for this hypothesis. At the same time, my finding is not compatible with Westera and Boleda's (2020) hypothesis that similarity should be negatively correlated with SI rate. Note that van Tiel et al. (2016) and Westera and Boleda (2020) both tested LSA scores as a predictor of scalar diversity but found that they were not a significant predictor. This discrepancy could be explained by my work testing a different scale set, but note also that as we will see in Section 4.4.7, LSA scores turn out not to remain a significant factor in explaining scalar diversity once we take more factors into account.

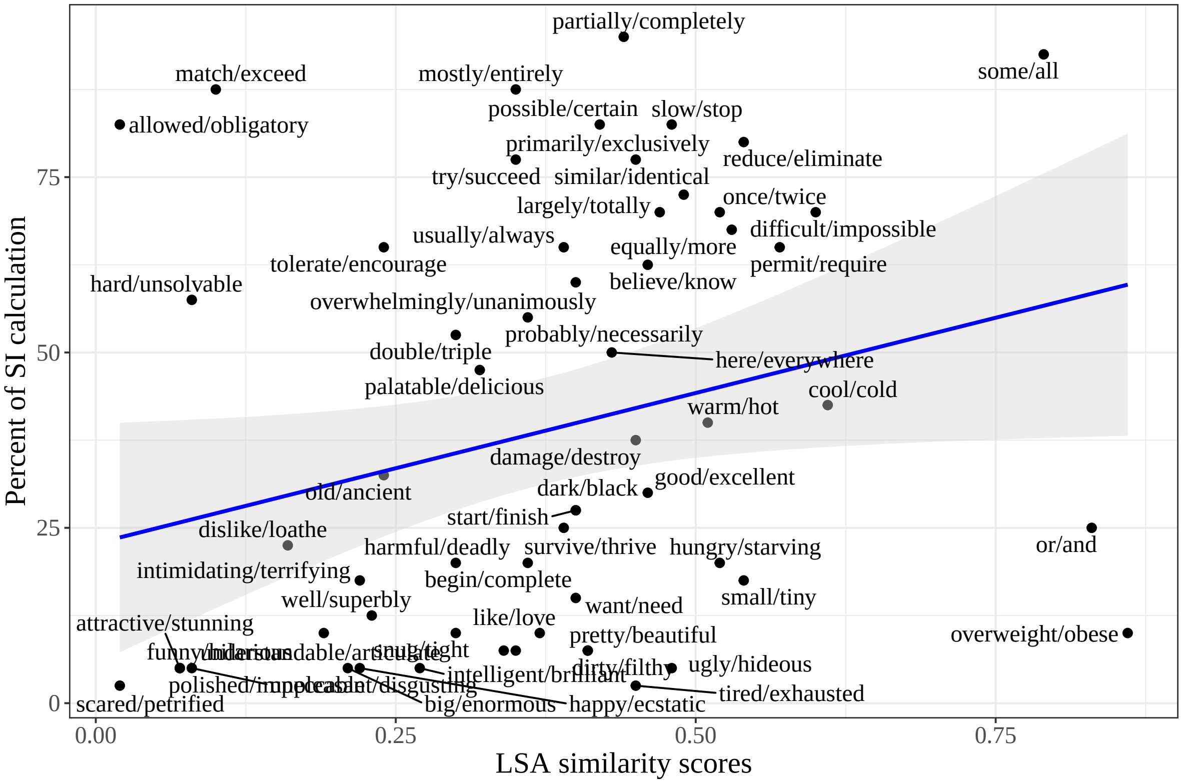


Figure 4.12: Results of the semantic similarity analysis: The x axis shows the LSA similarity scores. The y axis shows SI rate from Experiment 7.

GloVe

An alternative distributional semantic model is GloVe (Global Vectors; Pennington et al. 2014, which, like LSA, is count-based: it looks at word co-occurrences in a large corpus. In particular, I used spaCy word embeddings: 300-dimensional vectors that were trained through web documents for general-purpose tasks (en_core_web_lg, https://spacy.io/models). This measure of semantic similarity was not significantly correlated with SI rates (Pearson's correlation test: r=-0.12, p =.38). This is in line with Westera and Boleda (2020), who also tested the GloVe model, but found this kind of semantic similarity measure not to be a predictor of SI rates.

Word2Vec

Word2Vec belongs to the family of so-called "prediction-based" methods of distributional semantics. Unlike count-based methods, prediction-based ones start from random word vectors (e.g. weights in an artificial neural network) and incrementally update these vectors to better predict word-context occurrences. To get these similarity scores for the 120 scalar terms under investigation, Google's pre-trained Word2Vec model was used. Specifically, cosine similarity scores were computed between 300-dimensional Word2Vec embeddings, which were pre-trained on the Google News corpus (Mikolov et al., 2013). This was done via the interface provided by the gensim library for Python (Řehůřek and Sojka, 2010). Though there is a slight negative trend, such that the more similar two scalar terms are, the lower the SI rate, we do not find a significant correlation between Word2Vec similarity scores and SI rates (Pearson's correlation test: r=-0.198, p =.13). This also replicates Westera and Boleda (2020), who also did not find Word2Vec to be a significant predictor of SI rates.

In what follows, therefore, I use LSA scores as an index of semantic similarity, and do not analyze the Word2Vec and GloVe metrics further.

4.4.6 Experiment 10: Meaning of the negated strong scalar

In Experiment 10, I show that scalar diversity is (partially) explained by the meaning of the negated strong scalar term, as compared to the weak scalar. As in Experiment 9, my measure relies on experimentally collected degree estimates.

Hypothesis

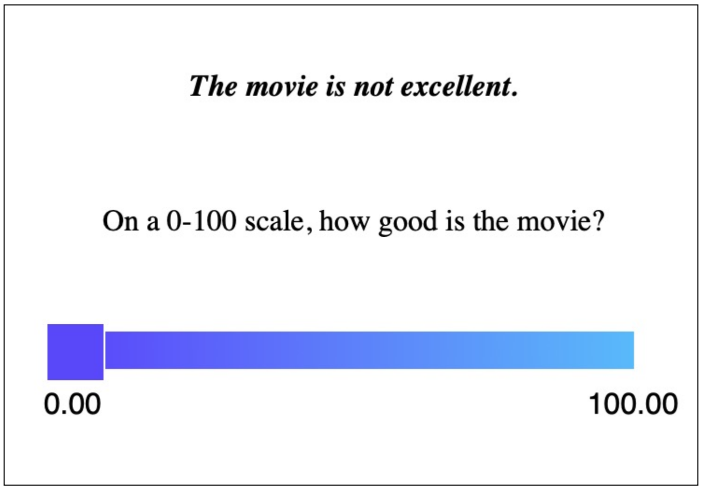
In this section so far, in line with much of the literature on scalar diversity, I focused on potential explanations for scalar diversity that had to do with the relationship between the weak and the strong scalar term. Experiment 10 takes a slightly different perspective, as it focuses on the meaning of the negated strong term (*not excellent*) as a predictor of the variation in SI rates. Let us first consider the inference task commonly used to test SI calculation, which I also used in Experiment 7. The inference task presents participants with an SI-triggering statement, such as *The movie is good*, and then poses the question: *Would you conclude from this that Mary thinks the movie is not excellent?*. (Neo)-Gricean accounts of SI calculation assume that hearers reason (only) about potential stronger alternatives to the weaker utterance they heard. But given the particulars of the inference task, it is conceivable that the meaning of the *negated* alternative (e.g., *not excellent*) also plays a role. In Experiment 10, I therefore probe what such negated stronger alternative statements mean, and what hearers therefore have in mind when answering the question of the inference task.

The specific hypothesis that I test is that the more similar the weak (*good*) and the negated strong (*not excellent*) term are, i.e. the smaller the difference between them on a degree scale, the higher the SI rate should be for that scale. Suppose, for instance, that *good* and *not excellent* are interpreted as describing two very different world states—that is, they are distant on the degree scale of goodness. In this case, it is implausible for a participant to conclude that a speaker meant *not excellent* when she uttered *good*. This can lead to a low rate of “Yes” responses in the inference task, which is then interpreted as a low SI rate.

Participants, task and procedure

31 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 7. Data from all 31 participants is reported below.

Experiment 10 had the same task and procedure as Experiment 9 —see Figure 4.13 for an example trial item. Here, I tested the negated strong term (in a between-participants design with Experiment 9). That is, participants saw the sentence *The movie is not excellent*, and then had to indicate on a 0-100 scale how good they thought the movie was.



The movie is not excellent.

On a 0-100 scale, how good is the movie?

Figure 4.13: Example experimental trial from Experiment 10: stronger scalar term

Prediction

My prediction for the results of Experiment 10 is that the smaller the difference between the degree estimates for the weak and the negated strong term, the higher the corresponding SI rate will be. In other words, I predict a negative correlation between the weak-not strong difference and SI rates.

Results and discussion

I conducted the same analyses as those reported in Experiment 9. Responses to negated strong terms were found to be significantly lower than to weak terms (Estimate=-33.59, Std. Error: 2.65, *t*=-12.65, *p*<0.001) —see Figure 4.14. That is, sentences such as *The movie is not excellent* received, on average, lower ratings on a 0-100 goodness scale than sentences such as *The movie is good*. I return to this finding below, in the discussion of negative strengthening.

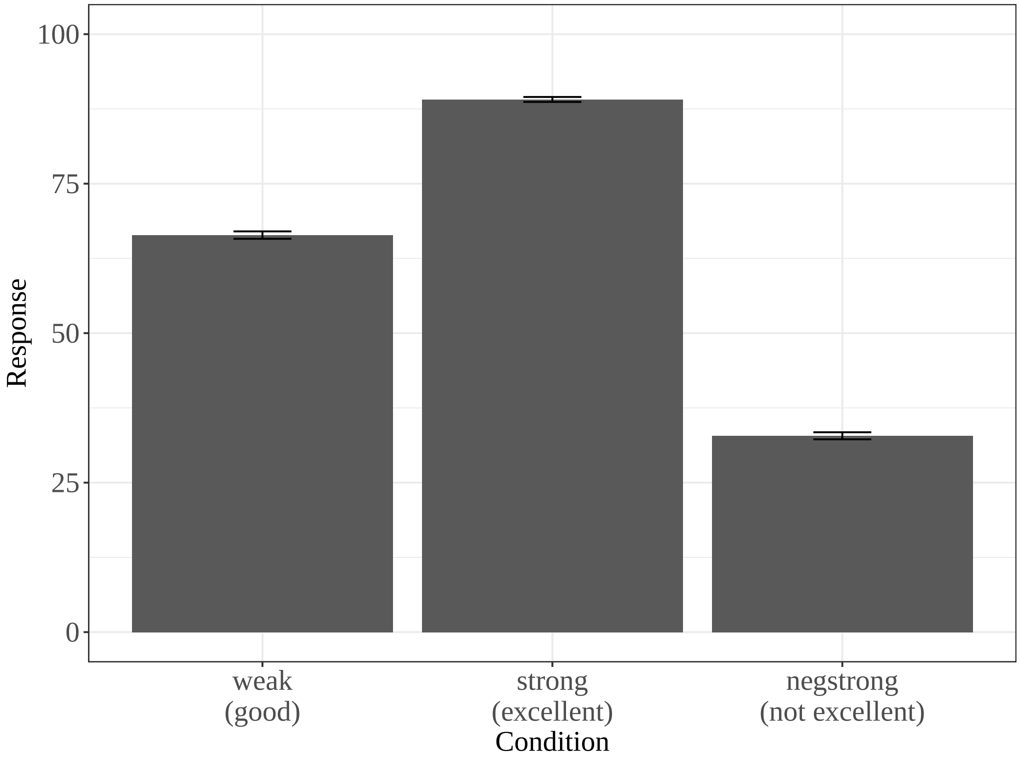


Figure 4.14: Results of Experiments 9 and 10: Average response on the degree estimate task (0-100 scale) for the two different conditions

To test the main prediction that the meaning of the negated strong term captures scalar diversity, I again calculated the absolute difference in means between the response to the weak term (from Experiment 9) and the response to the negated strong term (Experiment 10). For example, for the <*good, excellent*> scale, *The movie is good* received a response of 69.4 on the 0-100 scale, while *The movie is not excellent* received 31.5, resulting in a score of 37.9 —these are plotted on the *x* axis of Figure 4.15. There was a negative correlation between the results of Experiment 10 and 7 (Pearson's correlation test: r=-0.61, *p*<0.001); the more similar the world states that a weaker and negated stronger term are taken to describe, the higher the SI rate is for that scale. This constitutes evidence that the meaning of the negated stronger scalar plays a role in scalar diversity.

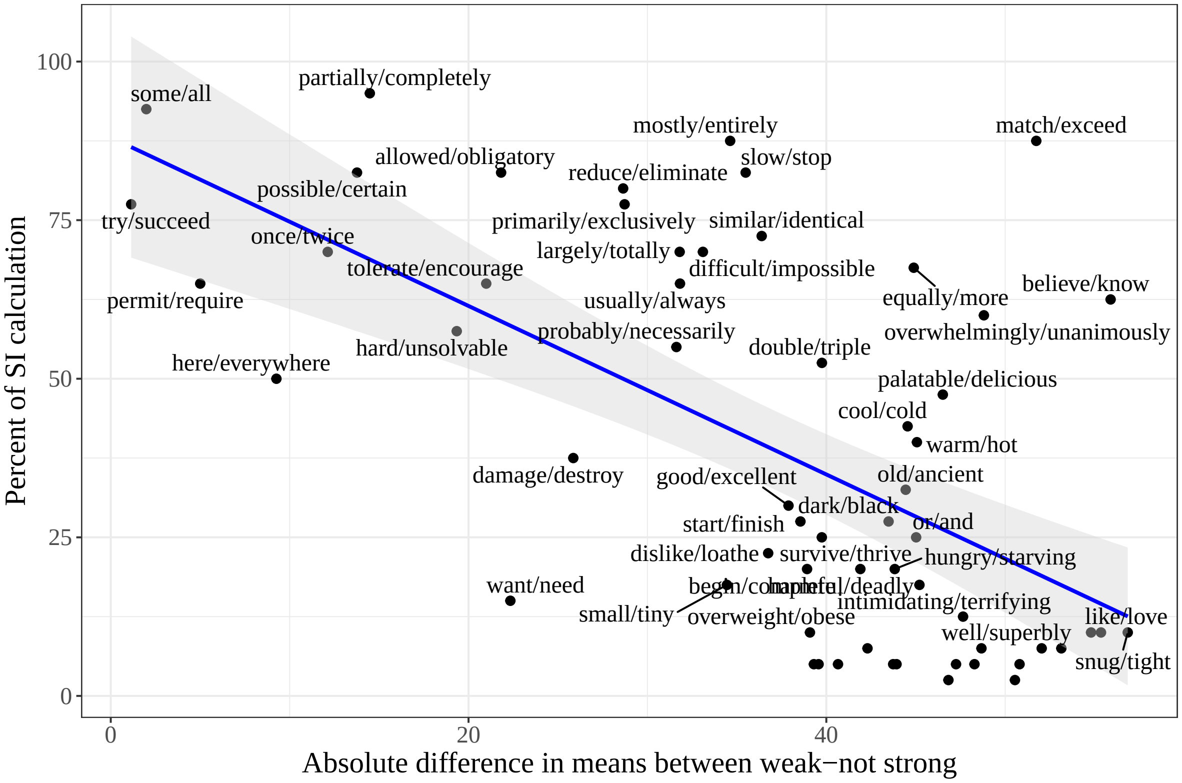


Figure 4.15: Results of Experiments 7 and 10: The *x* axis shows the meaning of the negated stronger term from Experiment 10. The *y* axis shows SI rate from Experiment 7.

To reiterate, the motivation for Experiment 10 was that the inference task commonly used to test SI calculation explicitly mentions the negated stronger term, raising the possibility that when participants choose not to endorse the conclusion that a speaker meant *not excellent* by uttering *good*, they do so because they perceive *not excellent* and *good* as meaning different things. My findings suggest that the meaning of the negated strong term, as measured by experimentally collected degree estimates, indeed captures some of the variation in SI rates that is observed across scales. This raises broader questions about whether the inference task is a good way to measure SI calculation. One limitation of the inference task in its current form is that it explicitly mentions, and therefore makes salient, the stronger alternative to a weaker scalar term (*Would you conclude... not **excellent**?*). Yet, scalar diversity emerges despite this potentially biasing nature of the task: we do not find that inference rates are uniformly high across scales, even though the stronger alternative is mentioned in the task question. The findings of Experiment 10 highlight a second potential problem with the inference task: namely, that it might introduce complications not only because it mentions stronger alternatives like *excellent*, but because it mentions *not excellent*, whose meaning I have shown to matter for SI calculation. For more recent discussion about task effects in experimental investigations of SI, see also Sun and Breheny (2021), who found that a task question like *Would you conclude from this that, according to Mary, not all of the questions are easy?* (similar to my Experiment 7) vs. one like *Would you conclude that, it could be that Mary thinks, all of the questions are easy?* produce different results.

As is reflected in the averages reported in Figure 4.14, the negated strong degree estimate was lower than the weak degree estimate for many lexical scales. This finding can be interpreted as negative strengthening, the pragmatic phenomenon where hearers take *John is not brilliant* to mean not only that John is less than brilliant (the sentence's literal meaning), but that he is less than intelligent, or that he is in fact stupid (Horn, 1989). As discussed in Section 4.1.2, Gotzner et al. (2018) experimentally tested propensity for negative strengthening across different scales: participants saw sentences such as *He is not brilliant* and were asked whether they can conclude that he is not intelligent. The authors found that "Yes" responses negatively correlated with SI rates and were able to predict scalar diversity. While negative strengthening is certainly relevant to the results of Experiment 10, there are a number of important respects in which my findings differ from Gotzner et al.'s. First, my collected data include scales that did not show negative strengthening, i.e. where the negated strong scalar term had a higher rating on the 0-100 degree scale scale than the weak scalar, suggesting that not all of Experiment 10's findings are attributable to negative strengthening. Second, though arguably tapping into similar pragmatic phenomena, negative strengthening is chiefly about *not brilliant* being interpreted as *not (even) intelligent*, while what I measured in this experiment is whether *not brilliant* is similar to *intelligent* in what world state it is taken to describe.

4.4.7 Combined analysis and variance explained

In this chapter, the following factors were shown to be correlated with SI rates:

- accessibility of the stronger alternative,
- distinctness of weaker and stronger scalar terms,
- boundedness,
- frequency of the stronger alternative,
- semantic similarity (only as indexed by LSA scores),
- meaning of the negated stronger scalar term

Next I report on statistical analyses conducted to check whether each of these factors remains a significant predictor when we take them all together, as well as how much of the observed variance each factor is able to capture. Using the lme4 package in R (Bates et al., 2015), I fit a logistic mixed effects regression model that predicted Response ("Yes" vs. "No") in Experiment 7's inference task as a function of the six factors listed above. The categorical predictor Boundedness was sum-coded before analysis ("bounded": -0.5 and "unbounded": 0.5).The model included random intercepts for participants. The model's estimates are shown in Table 4.4.

	Estimate	Std. Error	*z* value	*p* value
Intercept	1.59	0.31	5.11	
Accessibility	0.02	0	10.66	<0.001
Distinctness	-0.03	0	-9.41	<0.001
Boundedness	-1.04	0.12	-8.39	<0.001
Frequency	0.07	0.07	0.91	0.36
Semantic similarity	0.28	0.35	0.79	0.43
Meaning of negated strong term	-0.06	0	-11.96	<0.001

Table 4.4: Parameter estimates, standard errors, *z* values and *p* values from a logistic mixed effects regression model of the "Yes" vs. "No" responses in Experiment 7, predicted by the factors identified in this chapter

We can see that in this combined model, frequency and semantic similarity are no longer significant predictors of SI rates. Let us recall, however, that van Tiel et al. (2016) tested both of these factors (and Westera and Boleda (2020) tested semantic similarity), and found them not to be predictors for their own set of SI rate data. It seems plausible, therefore, that semantic similarity and frequency are indeed not relevant factors for the likelihood of SI calculation, and the correlations reported in Sections 4.4.4 and 4.4.5 merely reflect noise in the data.

But accessibility, distinctness, boundedness, and the meaning of the negated strong term continue to be significant factors. Therefore, to check how much of the variance these predictors explain (together and individually), I fit a new regression model, including only the four significant factors. The model was otherwise identical to the one detailed above. To check how much of the variance in the data is explained, I used the rsq package in R (Zhang, 2021) to compute R^2 values. I found that the model combining all four factors explained 28.2% of the variance in the data, with 24.6% coming from the fixed effects and 3.6% from the random effects. To test what proportion of the variance each factor explains, I checked how much the R^2 is reduced by fitting a model that removes that factor. That is, to calculate how much of the variance is explained by Accessibility, for instance, I fit a regression model only including the other three factors, and checked how the R^2 of that model compares to 24.6%[2]. (Here, I concentrate only on the variance explained by fixed effects.) Results are shown in Table 4.5.

	Estimate	Std. Error	z value	p value	R^2
Intercept	1.89	0.24	7.8		
Accessibility	0.02	0	11.65	<0.001	5.1%
Distinctness	-0.03	0	-9.32	<0.001	3.6%
Boundedness	-1.07	0.12	-8.7	<0.001	2.2%
Meaning of negated strong term	-0.06	0	-12.04	<0.001	5.8%

Table 4.5: Parameter estimates, standard errors, z values and p values from a logistic mixed effects regression model of the "Yes" vs. "No" responses in Experiment 7, predicted by the factors identified in this chapter

Altogether, this analysis finds that accessibility and the meaning of the negated strong term explain more of the variance than either distinctness or boundedness. But on its own, each factor can only capture less than 10% of the variance, suggesting that there is no single explanation for scalar diversity. Rather, the variance in SI rates across scales arises as a function of inter-

[2] A mixed effects model including only accessibility, distinctness and boundedness did not converge. I therefore fit a logistic regression model using the glm function in R to calculate the R^2 of the meaning of the negated strong term (5.8% in Table 4.5).

scale variation in multiple different properties. Future work should aim to synthesize all known predictors of scalar diversity reported in the literature (see Section 4.1.2), to give us an idea of how much of the total variance in SI rates across scales is now accounted for.

4.5 Summary of findings

This chapter introduced the phenomenon of scalar diversity, which complicates the picture that scalar inference arises simply as alternative-based reasoning, since such a story is (on its own) unable to directly predict the non-uniformity of scalar inference across different scales and alternatives. I also reported on a new set of lexical scales I collected via corpus work, which is larger than those in prior literature, and crucially it provides a better balance across different grammatical categories. I tested this new scale set in an inference task, which successfully replicated the scalar diversity phenomenon (Experiment 7), and I proposed quantifying the diversity using the information theoretic measure relative entropy.

Section 4.4 explored whether the variation in how likely a scalar inference is to arise can be derived from properties of the scalar alternatives themselves. We saw evidence that whether a scale is bounded, i.e., whether the stronger alternative denotes a fixed point (e.g., *all*) or not (e.g., *excellent*), predicts likelihood of SI calculation: bounded scales are more likely to lead to SI. I also showed that a production-based measure of how accessible a stronger scalar alternative is can capture scalar diversity: the more accessible the alternative (e.g., *all)*, the higher the SI rate (Experiment 8). The distinctness of the weak and strong scalar terms, as measured via degree estimates, was also found to be a predictor of SI rates: the more distinct the weaker term (*some*) and the stronger alternative (*all*) are, the higher the SI rate (Experiment 9). At the same time, it is not the case that all potentially relevant properties of alternatives are related to the likelihood of SI calculation: we saw that frequency of the stronger alternative or the semantic similarity of the weak and strong term do not play a role. In Experiment 10, I showed that the meaning of the negated stronger scalar term is also a significant predictor: the closer the meaning of *some* to *not all*, the higher the SI rate; in fact, this was the strongest predictor of the ones tested in this chapter.

This finding, however, might actually raise issues for the inference task itself that is typically used to measure SI calculation.

Finally, even though a number of relevant factors (most, but not all, related to the properties of alternatives) have been shown to predict scalar diversity, there is still unexplained variance, suggesting that scalar diversity remains a puzzle and a promising avenue for future work.

CHAPTER 5

SEMANTIC AND PRAGMATIC EFFECTS ON SCALAR DIVERSITY

Chapter 4 introduced the phenomenon of scalar diversity, and investigated potential factors that might explain the observed variation across scales[1]. In Chapter 5, I take a different perspective, and I look at how two factors unrelated to scalar diversity affect the likelihood and uniformity of calculating upper-bounded inferences (*some but not all*; *good but not excellent*). First, I manipulate the discourse context via an explicit question containing the stronger alternative (*all*, *excellent*), finding that such a supportive context both makes inference calculation more likely and reduces variability across scales (Section 5.2). This context manipulation is grounded in an elicitation experiment that establishes relevant QUDs across scales (Section 5.1). Following the investigation of context, I turn to the role of alternative exclusion. I find that overt exhaustification using the focus particle *only* (*only good*, *only some*), which encodes alternative exclusion semantically, also increases inference rates and leads to an even more robust reduction in scalar diversity (Section 5.3). Lastly, I show that when the two investigated factors align, i.e. there is pragmatic support from the context and a semantic cue to exclude alternatives, upper-bounded inferences are calculated at ceiling rates, and scalar diversity is eliminated (Section 5.4). Crucially, observations about whether variability is reduced are grounded in my proposal to more rigorously quantify scalar diversity than has been done in prior work, using the relative entropy measure introduced in Section 4.3.1.

5.1 Experiment 11: QUD elicitation

In Experiment 11, I further tackle the problem introduced in Chapter 3: that relatively little is known about the Questions Under Discussion that language users have in mind when presented with a given sentence. Similarly to Experiment 5 (Chapter 3), I develop and conduct an elicitation experiment that aims to track what kind of QUDs comprehenders have in mind when interpreting

[1] Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/xrezc/?view_only=cf7e4c36120b40b1b6d20bc0383a9e37

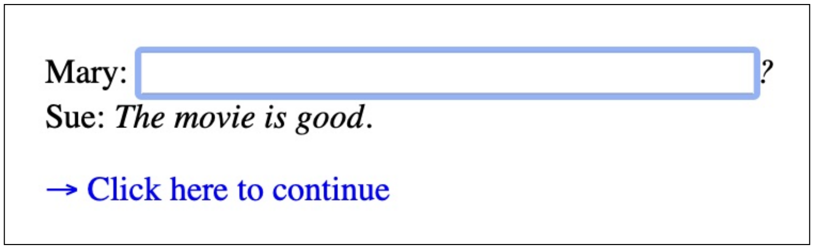
a potentially SI-triggering utterance —either on its literal, or on its SI-enriched meaning. Importantly, Experiment 11 investigates this question in the context of scalar diversity, testing 60 different lexical scales, not just <*some*, *all*> as in Experiment 5.

5.1.1 Participants and task

A total of 83 native speakers of American English participated in an online experiment on the Ibex platform (Drummond, 2007). Participants were recruited on Prolific and compensated $1.70 or $2.30 (depending on time commitment). Participant screening was identical to Experiment 7: one participant was removed due to failing attention checks. Data from 82 participants is reported below.

Experiment 11 employed an elicitation task. Participants were presented with a dialogue context where Sue uttered a potentially SI-triggering sentence, such as *The movie is good* (identical Experiment 7). This sentence was preceded by a question, uttered by Mary, which was left blank. Participants were instructed to guess what Mary's question was, by typing in their response. The experiment included a conclusion manipulation, which involved a neutral condition and an SI-biasing condition. In the neutral condition, participants only saw the two-sentence dialogue involving Mary's question (left blank) and Sue's answer (the SI-triggering sentence) —Figure 5.1 shows an example trial item. In the SI-biasing condition, a statement was presented in addition to the two-sentence dialogue, which spelled out to participants that Mary derived the SI from Sue's utterance. That is, for instance, if Sue's utterance was *The movie is good*, then the SI-biasing condition included a statement *Mary concludes that the movie is not excellent* —see Figure 5.2. Participants were instructed to guess what Mary's question was, given the rest of the dialogue and, in the case of the SI-biasing condition, the conclusion she ends up drawing.

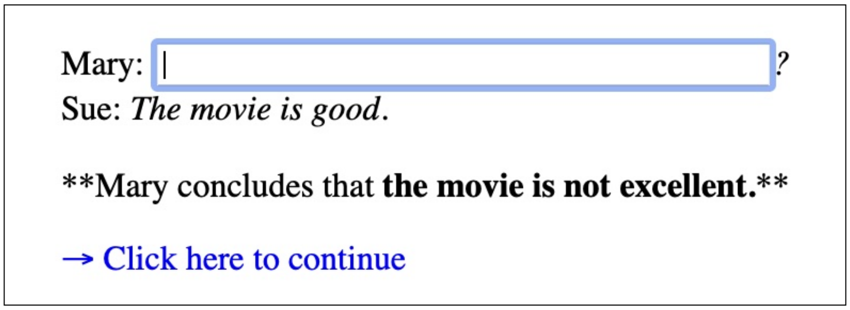
Items were modified for elicitation from the items of Experiment 7: a total of 60 critical items and 5 fillers were tested. The 60 critical items (i.e., 60 scales) were divided into two sets and tested separately. The conclusion manipulation (neutral vs. SI-biasing) was conducted between-participants. Therefore, a total of four sub-experiments were run: 21 participants were tested



Mary: ?
Sue: *The movie is good*.

→ Click here to continue

Figure 5.1: Example experimental trial from Experiment 11: neutral condition



Mary: |?
Sue: *The movie is good*.

Mary concludes that the movie is not excellent.

→ Click here to continue

Figure 5.2: Example experimental trial from Experiment 11: SI-biasing condition

on the first 30 scales (21 participants each for the neutral and the SI-biasing conditions), and 20 participants were tested (per condition) for the second 30 scales. This means that a total of 1230 data points were collected and analyzed for the SI-biasing condition, and another 1230 for the neutral condition.

5.1.2 Results and discussion

Results were coded in the following way. Each elicited question was categorized as one of the following seven categories. I first provide examples of the categories, based on three items: *The movie was good* → *not excellent*; *The boy is hungry* → *not starving*; *The student is intelligent* → *not brilliant*.

- Strong-scalar question: *Was the movie excellent?*; *Is the boy starving?*; *Is the student brilliant?*

- Weak-scalar question: *Do you think the movie is good?*; *Is the boy hungry?*; *Is the student intelligent?*

- Antonym question: *Is the student dull?*

- How-scalar question: *How good was the movie?*; *How intelligent is the student?*

- Negated scalar question: *The boy isn't hungry is he?*; *The student is not as bright as the others, huh?*

- Generic wh-question: *How was the movie?*; *What is the student like?*; *Is he a good student?*

- Miscellaneous: *So what did you think about that movie?*; *Why is the boy crying so much?*; *Is that student doing well in class?*

Having seen some representative examples, I now provide more details on what criteria were used to categorize the elicited questions. For the strong-scalar and weak-scalar questions, which are questions based on the stronger (e.g., *excellent*) or weaker (e.g., *good*) scalar term, attention was paid not just to whether the particular lexical item was present (e.g., *excellent*), but also to the meaning of the questions. In the majority of cases, the questions did contain the specific lexical items, but for illustration, I provide some examples of cases when they did not. For instance, a question containing *overjoyed* was categorized as strong-scalar for the <*happy*, *ecstatic*> scale, since it is a synonym of *ecstatic*; for the item *Ann's speech was polished* → *not impeccable*, a question with *good* was categorized as weak-scalar, i.e., synonymous with *polished* in this context; a question where everyone did something was taken to be synonymous with *unanimously* (strong-scalar); and a question such as *Which candidate is better?* was taken to be the same as *more* from *more skilled* (strong-scalar).

For something to be classified as an antonym, it needed to have the opposite polarity to the weaker and stronger scalar term, e.g., *cold* is negative polarity, while the <*warm*, *hot*> scale is positive. The how-scalar question category included questions that are not *how*-questions in their form, but whose meaning has to do with the degree to which something manifests the relevant scalar property. For instance, *What is the probability of success?* was classified as a how-scalar question for the item *Success is possible* → *not certain*. Additionally, questions that were a disjunction between the weaker and stronger scalar terms were also classified as how-scalar, e.g., *Was Stu's daughter petrified or just scared?* for the <*scared*, *petrified*> scale.

The "generic wh" and "miscellaneous" categories are characterized by what the discourse goal is in the dialogue, given Mary's question. In the strong-scalar, weak-scalar, antonym, how-scalar and negated scalar questions, the discourse goal has to do with what degree of the relevant property (e.g., goodness, intelligence, or quantity) is held —in other words, a question was coded as belonging to one of these categories if it established a QUD about scales or degrees. In the generic wh and miscellaneous questions, on the other hand, the discourse goal is something different, e.g., to learn a reason for something (*Why is the boy crying so much?*), or learn about a person (*What is the student like?*), or learn someone's identity (*Who got the highest score between Bill and Al?*). The difference between these two categories (generic wh and miscellaneous) is ultimately not of interest to us.

Table 5.1 shows the results of Experiment 1: both the raw counts and frequencies that resulted from the coding of elicited questions.

	SI-biasing		**Neutral**	
	Percentage	Count	Percentage	Count
Strong-scalar	28.1%	346	0.7%	9
Weak-scalar	8.9%	109	9.8%	120
Antonym	4.3%	53	3.5%	43
How-scalar	12.4%	152	4.2%	52
Negated scalar	3.3%	41	0.2%	3
Generic wh	21.5%	264	22.8%	281
Miscellaneous	21.5%	265	58.7%	722
TOTAL	100%	1230	100%	1230

Table 5.1: Experiment 11 results: frequencies of elicited question types

Comparing the two conditions, neutral and SI-biasing, we see that the SI-biasing condition resulted in more strong-scalar questions, as well as more how-scalar questions —though how-scalar questions remain overall relatively infrequent. The neutral condition, on the other hand, resulted in more miscellaneous questions. A number of conclusions emerge from these results of the QUD elicitation experiment. First, we can take these findings as evidence from language production for the formerly comprehension-based finding that QUDs can encourage or discourage SI calculation (for this finding in the context of scalar diversity, see Cummins and Rohde 2015; Ronai and Xiang 2021, and Experiment 12 below). Second, it seems that questions that bias toward the SI-enriched interpretation of the sentence need to make reference to the stronger scalar term; no other strongly SI-biasing questions emerged, other than the strong-scalar questions. This kind of question indeed provides the basis for the QUD manipulation in comprehension experiments such as Ronai and Xiang's (2021) Experiment 3 and Experiment 12 of this chapter. Lastly, the elicited data did not reveal a by-item effect: the 60 different scales did not differ noticeably from each other in the types of questions they elicited and whether they elicited more of one type (e.g., strong-scalar vs. weak-scalar). This finding is broadly in line with how a QUD manipulation in a comprehension experiment does not eliminate the variation across scales, i.e., scalar diversity (see Ronai and Xiang 2021 and Experiment 12).

5.2 Experiment 12: Discourse context manipulation

Much of the experimental pragmatics literature, including on scalar diversity, and including Experiment 7 in this dissertation, presents stimulus sentences in the absence of any context. But it is well known that properties of the discourse context, formalized e.g. as Question Under Discussion (QUD, Roberts 1996/2012), can make SI calculation more or less likely (Van Kuppevelt, 1996). Indeed, experimental work has confirmed this modulating role of context not only for the <*some*, *all*> scale (Degen and Tanenhaus 2014; Yang et al. 2018; Zondervan et al. 2008; Chapter 3 of this dissertation), but also for a variety of different lexical scales (Cummins and Rohde, 2015; Ronai and Xiang, 2021).

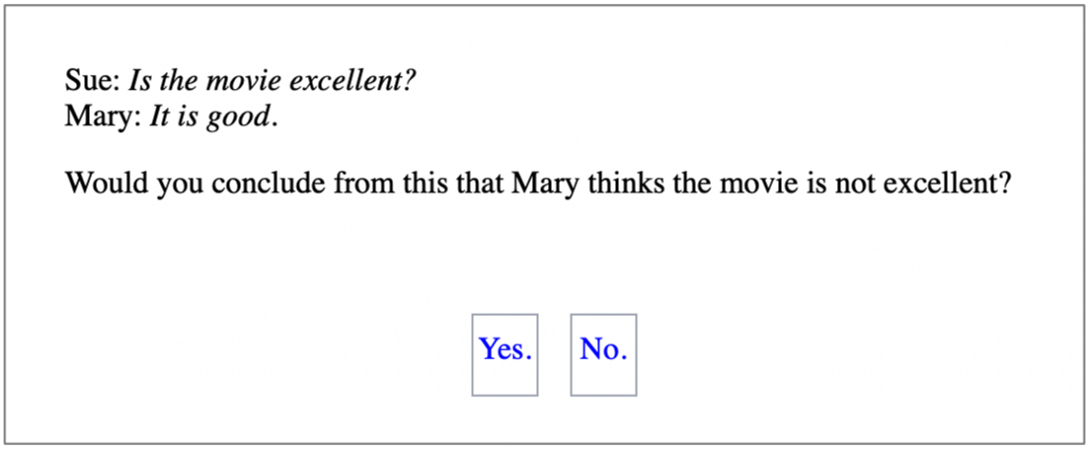
In Experiment 12, I operationalize discourse context as explicit questions, and investigate the effect of such a manipulation on the likelihood of SI calculation, as well as on the observed variation across scales, that is, scalar diversity.

5.2.1 Participants and task

81 native speakers of American English participated in an online experiment, administered on the Ibex (Drummond, 2007) and PCIbex (Zehr and Schwarz, 2018) platforms. Participant recruitment, screening, and compensation was identical to Experiment 7. Data from all 81 participants is reported below.

Experiment 12 employed the same task as Experiment 7, but the potentially SI-triggering sentences (uttered by Mary) were now embedded in a dialogue context. Specifically, the SI-triggering sentences were either preceded by a polar question that contained the stronger scalar ("strong QUD" condition), or by a polar question that contained the weaker scalar term ("weak QUD" condition). For the <*good*, *excellent*> scale, for instance, the manipulation included the question *Is the movie excellent?*–see Figure 5.3–, or the question *Is the movie good?*. The question manipulation (strong vs. weak QUD) was administered within participants in a Latin Square design.

Mary's answers were modified to ensure dialogue coherence, e.g., *The movie is good* was changed to *It is good*. Otherwise, Experiment 12's materials and procedure were identical to Experiment 7.



Sue: *Is the movie excellent?*
Mary: *It is good*.

Would you conclude from this that Mary thinks the movie is not excellent?

Figure 5.3: Example experimental trial from Experiment 12

5.2.2 *Hypothesis and predictions*

Standard semantic treatments of questions take them to partition a set of possible worlds into cells denoting their possible answers (Hamblin, 1976; Groenendijk and Stokhof, 1984). The question *Is the movie excellent?*, then, partitions the Common Ground based on the stronger alternative *excellent*: in one cell are all the worlds where the movie is excellent, and in the other cell, all the worlds where the movie is not excellent. An answer, in turn, is taken to be congruent with (or "a good answer to") a question if it determines which cell contains the actual world (Hulsey et al., 2004). Consider the two readings (literal and SI-enriched) of a potentially SI-triggering sentence in this light:

(1) The movie is good.

a. The movie is at least good. literal

b. The movie is good, but not excellent. SI

Given the question *Is the movie excellent?*, the SI-enriched meaning in (1-b) is a congruent answer, because it entails the "not excellent" cell of the partition, and eliminates the "excellent" cell. The literal meaning in (1-a), on the other hand, does not entail either cell, and it therefore does not directly bear on the question. Therefore, only on its SI-enriched meaning does *The movie is good* constitute a congruent answer.

The picture is different, however, in the weak QUD condition. A polar question like *Is the movie good?* partitions the Common Ground such that in one cell are all the worlds where the movie is good, and in the other cell, all the worlds where the movie is not good. An answer of *The movie is good* constitutes a good answer here no matter whether it gets enriched to mean *not excellent*, since it entails the "good" cell of the partition in either case.

Given this, I hypothesize that the strong QUD manipulation in Experiment 12 will encourage SI calculation; that is, participants will calculate SIs in order to make Mary's answers congruent with Sue's questions. Consequently, I first predict that SI rates will increase in the strong QUD condition, as compared to the baseline Experiment 7, which included no context. I also predict that the strong QUD condition will produce higher SI rates than the weak QUD condition. In fact, in the strong QUD condition, inference rates could increase to ceiling (100%), since without SI calculation, the dialogue participants are presented with would not be congruent. Second, I predict that as inference rates increase across the board for all scales, variation across scales (scalar diversity) will be reduced. In the weak QUD condition, on the other hand, there is no reason to predict a rise in SI rates as compared to Experiment 7, since Mary's answer will be congruent no matter whether it receives an SI interpretation.

5.2.3 Results and discussion

Figure 5.4 shows the results of Experiment 12 (second facet: "Weak QUD" and third facet: "Strong QUD"), along with the result of Experiment 7 (first facet: "SI"). To compare the rates of inference calculation in the weak vs. strong QUD conditions, I 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 Condition. It included the maximal random effects structure supported by the data (Barr et al., 2013): random intercepts for participants and random slopes and intercepts for items. The fixed effects predictor Condition (weak QUD vs. strong QUD) was sum-coded before analysis (weak QUD: -0.5 and strong QUD: 0.5). This analysis revealed that the strong QUD condition led to higher inference rates than the weak QUD condition (Estimate=1.67, SE=0.1, z=16.15, p <0.001).

In the next analysis, I compared the rates of inference calculation in Experiment 12 (in either the weak or the strong QUD condition) to the inference rates from Experiment 7. Two separate models were fit, one comparing strong QUD to Experiment 7, and another comparing weak QUD to Experiment 7. These logistic mixed effects regression models predicted Response ("Yes" vs. "No") as a function of Experiment. The maximal random effects structure supported by the data (Barr et al., 2013) included random intercepts for participants and random slopes and intercepts for items for the comparison of strong QUD vs. Experiment 7, and random intercepts and slopes

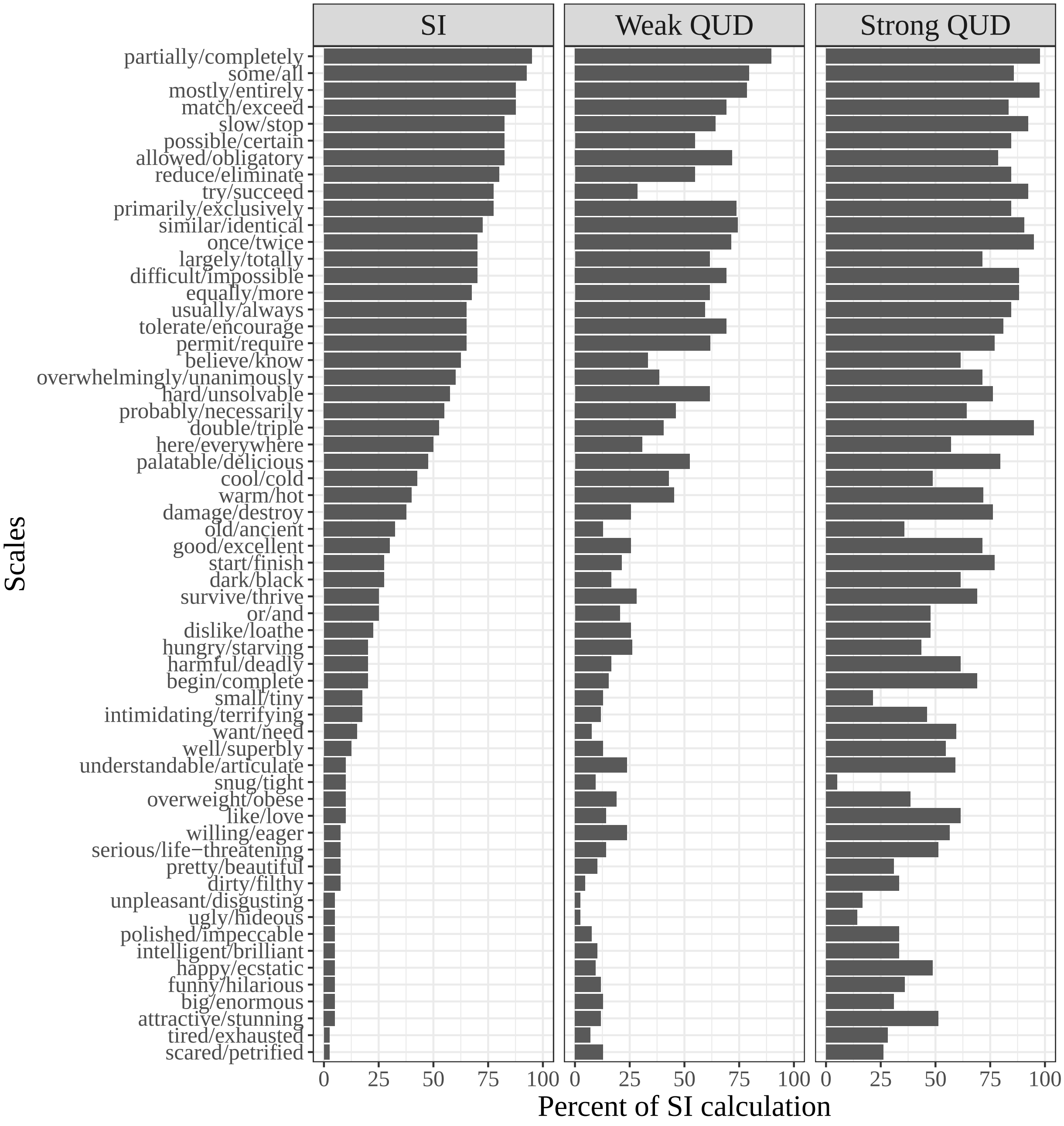


Figure 5.4: SI rate for 60 different scales. Experiments 7 (SI) and 12 (Weak QUD, Strong QUD) are shown on the three facets of the plot.

for both participants and items for the comparison of weak QUD vs. Experiment 7. The fixed effects predictor Experiment (7 vs. 12) was sum-coded before analysis (Experiment 7: -0.5 and Experiment 12: 0.5). This analysis revealed an overall increase in inferences rates in Experiment 12's strong QUD condition, as compared with Experiment 7 (Estimate=1.49, SE=0.22, z=6.56, p <0.001). In the weak QUD condition, however, inference rates were not statistically different from those in Experiment 7 (Estimate=-0.25, SE=0.18, z=-1.38, p =0.17).

Overall, we find that a supportive discourse context (i.e., strong QUD) made participants significantly more likely to calculate inferences, as compared to either a no-context situation (Experiment 7), or context with a weak QUD. These findings replicate Ronai and Xiang's (2021) Experiment 2 on a different, larger set of scales.

To check the effect of the discourse context manipulation on the variation in SI rates across scales, we can turn to our measure of relative entropy. The SI rates in the weak QUD condition of Experiment 12 resulted in a relative entropy of 0.378, while the strong QUD condition resulted in a relative entropy of 0.123. Recall that lower numbers represent more uniformity: if all scales led to SI at the same rate, relative entropy would be 0, but the relative entropy of the baseline Experiment 7 (without context) was 0.466. What we find, then, is that a supportive context reduced the variation in SI rates: relative entropy is lower in the strong QUD condition of Experiment 12 than in Experiment 7, i.e., there is less scalar diversity. The weak QUD condition did also lead to a slight reduction in diversity compared to Experiment 7, but much less so than the strong QUD condition. Altogether, in line with my predictions, an explicit question based on the stronger scalar term both increased SI rates across the board and reduced the variation across scales —but a question based on the weaker scalar term did not have the same effect.

At the same time, even with the strong QUD, we did not find a ceiling effect in SI rates, nor did we find uniformity across scales. This presents a puzzle. As discussed, a question such as *Is the movie excellent?* partitions the Common Ground into possible worlds where the movie is excellent vs. possible worlds where the movie is not excellent. Given this, Mary's utterance only constitutes a felicitous contribution to the discourse on its SI-enriched meaning (*good but not* *excellent*), because only on this meaning does it entail one of the cells of the partition. Since this is the same for all scales tested, we would expect equally high SI calculation everywhere.

I propose the following possible reason for why the predicted uniformity does not obtain: there are in fact three different possible pragmatic meanings that can be attributed to Mary's utterance in the dialogue context, which I detail below in (2-a)-(2-c).

(2) Sue: Is the movie excellent?
Mary: It is good.

a. It is good (but not excellent). SI

b. (Well,) it's good. ignorance

c. (Yes,) it's good. good $\approx$ excellent

Example (2-a) is the standard scalar inference meaning, which was the one I intended for participants to arrive at in Experiment 12. It is also possible, though, that (some) participants assigned to Mary's answer the meaning in (2-b), which is communicating ignorance about the stronger alternative; on this reading, Mary's answer conveys not that the movie is not excellent, but that Mary does not know whether it is excellent. Lastly, (2-c) shows a third possibility, where *good* is used as a synonym for *excellent*—Mary is in fact giving an affirmative answer to Sue's question. Using a weaker scalar term as a synonym for a stronger alternative may be related to the semantic distance between or distinctness of the two scalar terms, which has been shown to independently correlate with SI rates. As van Tiel et al. (2016) and Experiment 9 in Chapter 4 have demonstrated, the less distant or distinct the two scalar terms are, the less likely the SI is in a no-context situation (like Experiment 7). Additionally, the less distinct they are, the more it is possible to interpret the weaker term as a synonym for the stronger alternative, as in (2-c), and the lower the SI calculation rate stays even with a biasing context (like the strong QUD condition of Experiment 12).

It is also possible to analyze the interpretation in (2-c) as an R-implicature (Horn, 1984). (3) shows a classic example of an R-implicature, where a statement of (3-a) implicates (3-b).

(3) a. I need a drink.

b. I need an alcoholic drink.

In R-implicature, a generic form (*drink*) takes on a more specific meaning (*alcoholic drink*). In other words, while scalar inferences introduce upper-bounded meanings (*good* means *not more than good*), R-implicatures are lower-bounding: what is said is the lower limit of what is actually the case. Arguably, taking *The movie is good* to affirm that the movie is excellent is an instance of such an implicature.

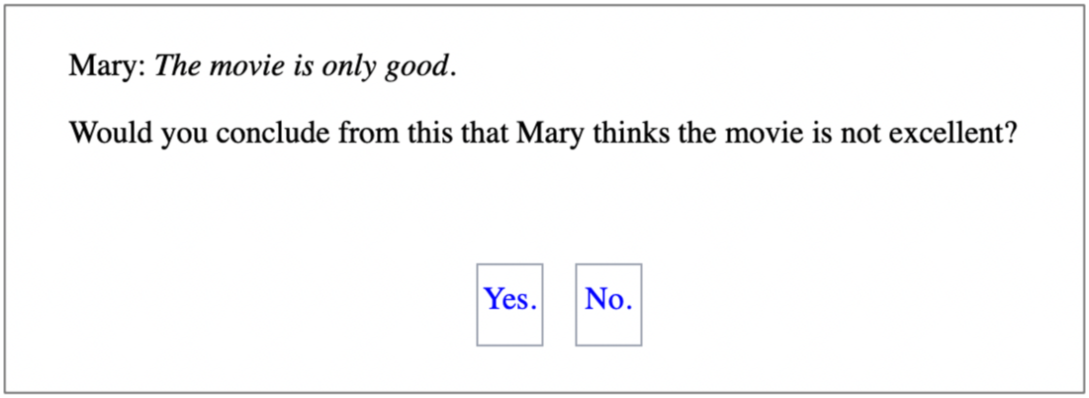
Even though the experimental manipulation intended for participants to arrive at the meaning in (2-a), it must be noted that (2-a) and (2-c) both represent congruent answers: (2-a) addresses the question by entailing "not excellent", while (2-c) addresses it by entailing "excellent" —assuming that (2-a) and (2-c) are interpreted on their respective enriched meanings. But only if a participant had (2-a) in mind did they answer "Yes" in the inference task; with either (2-b) or (2-c), they answered "No". Crucially, all three different readings for Mary's *It is good* answer should correspond to different prosodic contours; in particular, the meaning in (2-b) would correspond to the rise-fall-rise contour, see Ward and Hirschberg (1985) and Constant (2012), but also de Marneffe and Tonhauser (2019) and Göbel and Wagner (2022). Directly manipulating prosody using audio stimuli to tease apart these possible readings is therefore a promising area for future work.

5.3 Experiment 13: Focus manipulation

While Experiment 12 tested a pragmatic manipulation, Experiment 13 uses a semantic one: I investigate the effect of focus (signalled by the particle *only*) on the likelihood and diversity of inference calculation.

5.3.1 Participants and task

41 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment, screening, and compensation was identical to Experiment 7.


Mary: *The movie is only good*.

Would you conclude from this that Mary thinks the movie is not excellent?

Figure 5.5: Example experimental trial from Experiment 13

Data from 40 participants is reported below.

Experiment 13 employed the same inference task as the previous two experiments. This time, the additional manipulation conducted was to include the focus particle *only* in the SI-triggering statement. That is, Mary's utterance was e.g., *The movie is only excellent*—see Figure 5.5 for an example trial. Other than this, the materials and procedure were identical to Experiment 7.

5.3.2 Hypothesis and predictions

The focus operator *only* semantically excludes alternatives to the focused element (Rooth, 1985, 1992). That is, the sentence *The movie is only good* conveys that the movie is good (=positive component), and crucially also that the movie does not hold other properties from some contextually determined set of alternatives (=negative component). On a standard Roothian semantics (Rooth, 1985, 1992), the particle *only* associates with the element in focus (*good*) and expresses that, among the set of possible alternatives {wonderful, excellent, ...}, the movie does not hold any other property. The focus operator thereby excludes the alternatives to the focused element.

Unlike in previous experiments, where the exclusion of the stronger alternative (e.g., *excellent*) was a cancellable pragmatic inference, in Experiment 13, alternative exclusion is an entailment. Based on this, I predict that comprehenders will exclude alternatives to the focused element, and consequently that the rates of upper-bounded inference calculation will increase. Again, inference rates could possibly increase to ceiling (100%), given that *The movie is only good* encodes the exclusion of alternatives to *good* in a non-cancellable way. As a consequence of inference rates increasing across the board, I also predict that the variation (diversity) observed across lexical scales will be reduced.

5.3.3 Results and discussion

Figure 5.6 shows the results of Experiment 13 (third facet: "Only"), along with the results of Experiments 7 and 12. To compare Experiment 13's results to that of the baseline Experiment 1, I conducted the same statistical analysis as the one reported in Section 5.2.3. This analysis revealed that Experiment 13 also led to significantly higher rates of inference calculation than Experiment 7 (Estimate=1.86, SE=0.27, z=6.96, p <0.001)—participants were more likely to calculate upper-bounded inferences in the presence of overt exhaustification with *only*. Turning now to our measure of the "diversity" of inference rates, Experiment 13 led to a relative entropy value of 0.046. Compared to the previous experiments (see Table 5.2), we see a more substantial reduction in variation across scales; scalar diversity was lessened more with the focus manipulation than with the discourse context manipulation.

	Manipulation	Relative entropy
Experiment 7	Baseline scalar diversity	0.466
Experiment 12	Weak QUD	0.378
Experiment 12	Strong QUD	0.123
Experiment 13	Exhaustification with *only*	0.046

Table 5.2: Relative entropy results by experiment: Experiments 7, 12 and 13

In line with the predictions, then, the focus manipulation made upper-bounded inference calculation (*some but not all*, *good but not excellent*) more likely, and it also reduced scalar diversity. Additionally, *only* had a bigger effect than the strong QUD in Experiment 12, both in how much it increased inference rates and how much it reduced diversity. This finding makes sense in that, though both manipulations are predicted to increase inference rates, the question manipulation is fundamentally a pragmatic one. With the focus manipulation, the cue to exclude alternatives is encoded in the semantics, i.e., the grammar. In contrast, the pragmatic manipulation encourages

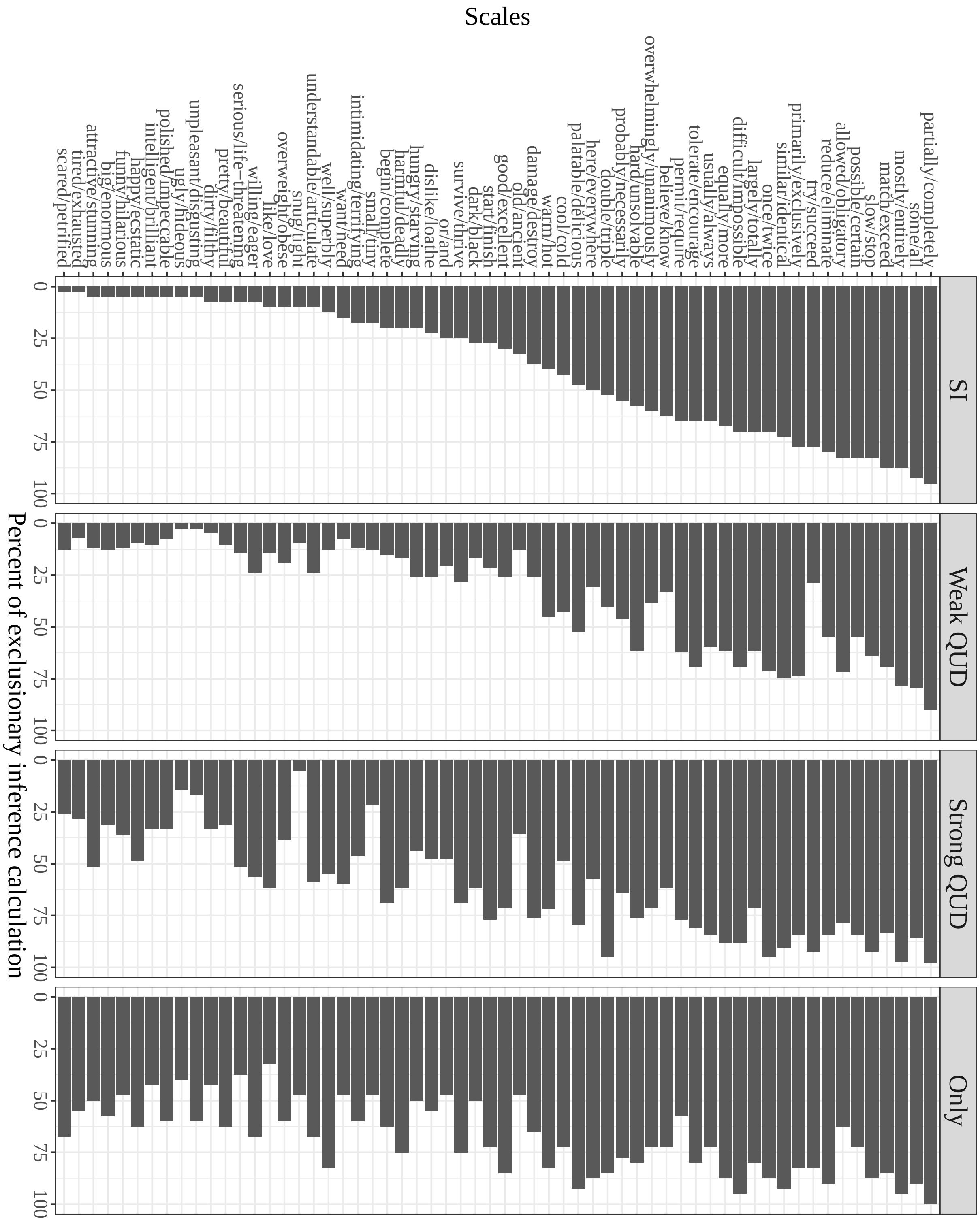


Figure 5.6: SI rate for 60 different scales. Experiments 7 (SI), 12 (Weak QUD, Strong QUD) and 13 (Only) are shown on the four facets of the plot.

inference calculation to ensure dialogue coherence, which is a more violable constraint than grammatically encoded alternative exclusion.

However, as is evident from Figure 5.6, we do not have ceiling-level inference rates: it is not the case that encoding alternative exclusion in the semantics always led participants to answer "Yes" in the inference task for all scales. Additionally, there still remains variability across the different scales. This finding might have implications for theories of scalar inference that derive SI-enriched meanings as part of the compositional semantics (Chierchia, 2004; Chierchia et al., 2012). On such theories, *Mary ate all of the deep dish* gets its *Mary didn't eat all of the deep dish* interpretation via exhaustification; it is assumed that such sentences contain a covert operator that serves the same function as its overt counterpart *only*. But as Experiment 13 demonstrates, even when we do in fact have an overt *only*, the calculation of exclusionary inference is still not uniformly high. This might pose a challenge to theories that encode SI in the grammar: why would SI be encoded in the grammar, captured via a silent exhaustification operator, if we see that even in the presence of an overt exhaustification operator, the relevant upper-bounded meanings are not always derived? Though of course, it should be possible to supplement grammatical accounts of SI with factors predicting the variable robustness of SI derivation across scales (such as the factors in Chapter 4). Additionally, as I will argue below, the findings with (overt) *only* can also be given an explanation.

I propose two potential (related) explanations for the observed lack of uniformity and ceiling-level inference rates. First, *only* is ambiguous between its so-called rank-order reading and its complement-exclusion reading (terminology from Coppock and Beaver 2013, original observation from Horn 1969). The rank-order reading concerns the placement of *good* on a scale where elements are ordered by rank. This reading of *only* can be paraphrased as *no more than*: *The movie is only good* means that the movie is no more than good. A complement-exclusion reading, on the other hand, excludes all alternatives to the focused element, including those that are not ordered with respect to the focused element on some scale. This reading of *only* can be paraphrased as *nothing other than*: *The movie is only good* means that the movie is nothing other than good. One example where the two different readings clearly come apart is the sentence *She's only an assistant* *professor* —on the rank-order reading, this means that she is not an associate or a full professor, while on the complement-exclusion reading, it means that she is not, for instance, a librarian.

It must be noted that there are uniform semantic treatments of rank-order and complement-exclusion readings. Coppock and Beaver (2013) (see also Beaver and Clark 2008) give a uniformly scalar treatment to both readings, arguing that the negative component is in fact always expressible with *no more than*. On a scalar analysis of rank-order readings, the scale of alternatives corresponds to a rank-ordering (Horn, 2000b). For *The movie is good*, this scale could, for instance, be <*okay*, *good*, *excellent*> —see Figure 5.7. Here, boldface represents the positive component of *only* (*The movie is (at least) good*), while alternatives ruled out by the negative component (*The movie is no more than good*) are struck out.

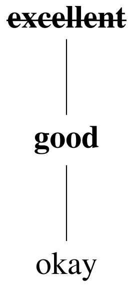


Figure 5.7: Scalar analysis of the rank-order reading of *only*, adapted from Coppock and Beaver 2013, ex. 29

Perhaps more interestingly, the complement-exclusion reading can also be obtained in a scalar framework: for this, alternatives are ranked as a boolean lattice —see Figure 5.8. The nodes here are shorthand for possible alternative answers to a question such as *What properties does the movie have?*: *The movie is good*, *The movie is funny*, *The movie is good and funny*, etc. In bold are the alternatives ruled in by the positive component of *The movie is only good*, while alternatives that are ruled out by the negative component of *only* are struck out.

Importantly for the interpretation of the experimental findings, even if we give a uniform semantic treatment to the two possible readings, they still represent two distinct readings of *only* and correspond to the ruling out of (potentially) distinct alternatives. On the rank-order reading, the stronger alternative *excellent* must be excluded; if participants were all assigning this meaning to

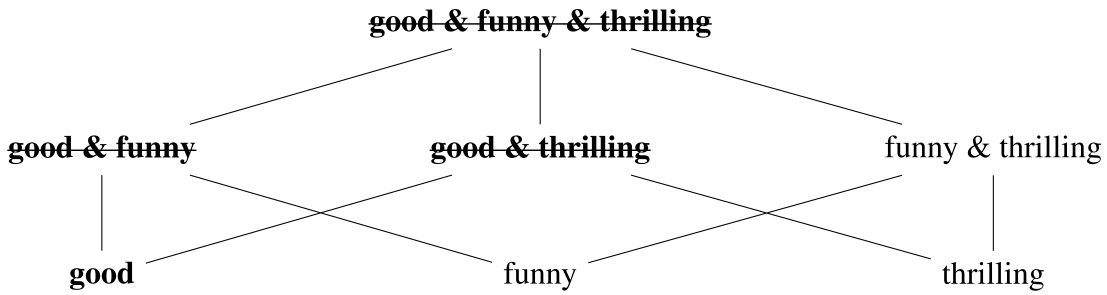


Figure 5.8: Scalar analysis of the complement-exclusion reading of *only*, adapted from Coppock and Beaver 2013, ex. 27

only, we should be seeing ceiling level "Yes" responses. On the complement-exclusion reading, however, it is possible to interpret Mary's utterance in the experiment as communicating that the movie is good, but not funny or thrilling, etc. Excluding such alternatives leaves open the possibility that the movie is in fact excellent, leading participants to respond "No" in the inference task.

Relatedly, in Experiment 13, stimulus sentences appeared without any context. It is therefore possible that participants had different contexts in mind. Compare, for instance, (4) and (5).

(4) Sue: Is the movie excellent?
Mary: It is only good.

(5) Sue: What's the movie like?
Mary: It is only good.

As discussed in relation to Experiment 12, a context like (4) encourages enriching *good* to *good but not excellent* in order to yield a congruent answer —even in the absence of *only*. Given such a context, then, *only* is most naturally interpreted as excluding this alternative *excellent*. If a participant supposed such a context, they would arrive at the upper-bounded *good but not excellent* inference, and answer "Yes" in the experimental task. But if they supposed a context like the one in (5), the alternatives that are to be excluded could be any property that a movie can have. Therefore, on this interpretation, participants could have concluded that the movie is only good, but not funny, thrilling, scary, etc., ultimately answering “No” to the task question about the movie not being excellent.

More generally, focus and questions (including QUD) are closely related (Rooth, 1985, 1992; Beaver and Clark, 2008, i.a.). For instance, an answer to a wh-question has to have a congruent focus structure: the position of the focused item in the answer has to correspond structurally to the position of the wh-word in the question. For this reason, B’s answer to A in the following dialogue is infelicitous with focus on the verb *likes*, but felicitous with focus on *dancing* (where focus is marked with capital letters).

(6) A: What does Mary like?
B: #Mary LIKES dancing.
B$'$: Mary likes DANCING.

It is no surprise, then, that in Experiment 13, where context was left unspecified, there remained uncertainty over the specific interpretation of the *only*-containing sentences.

To summarize, while overt exhaustification with the focus particle *only* encodes alternative exclusion in the semantics, it is possible that participants interpreted Mary’s utterance as excluding some alternative(s) other than the one the experiment tested (i.e., for *good*, something other than *excellent*). I argue that this might be the reason inferences rates were not at ceiling in Experiment 13, and why some of the inter-scale variation still remained[2].

An interesting avenue for future research is to test exclusives other than *only*, which vary in whether they allow both rank-order and complement-exclusion readings (Coppock and Beaver, 2013). For instance, *exclusively*, *solely* and *purely* only have the complement-exclusion reading. Consider the below dialogue, in which A’s question brings to mind stronger rank-order alternatives to *assistant professor*, e.g., *associate professor (with tenure)*.

[2] For a small minority of items, it is also possible that there was ambiguity not in the identity of the alternative, but in the focus associate itself. For example, the sentence *The princess only likes dancing* (intended SI: *She doesn’t love dancing*) could also be interpreted such that *only* associates not with *like*, but with *dancing*, leading to the inference that the princess does not like activities other than dancing.

(7) A: Has Mary received tenure?

B1:#She's exclusively/solely/purely an assistant professor.

B2: She's only an assistant professor.

(adapted from Coppock and Beaver, 2013, ex. 170)

Given the context of such a question, B1's answer is infelicitous. This is because the expectation is that B1's answer would communicate that the highest-ranked relevant property that Mary holds is assistant professor, but the exclusives *exclusively*/*solely*/*purely* do not allow a rank-order reading. In contrast, B2's answer is felicitous, since *only* licenses the required rank-order reading.

On the other hand, the exclusive *mere* only has a rank-order reading: the below example means that Google is *no more than a conduit*, i.e., not a collaborator in the deception (which would be a higher-ranked alternative). The sentence does not mean that the only property Google holds is being a conduit. For instance, it is still true that Google is a search engine.

(8) At trial, Nicholas J held that Google was a 'mere conduit' for the advertiser's misleading or deceptive representations. (Coppock and Beaver, 2013, ex. 33)

Somewhat similarly to *mere*, *merely* also prefers the rank-order reading, as demonstrated by the infelicity of *merely* in the below sentence.

(9) I (only/#merely) like [Apple computers$_F$]. (Coppock and Beaver, 2013, ex. 177)

Though with *merely*, complement-exclusion readings are not completely out, as can be seen in:

(10) An epicurean is someone who likes merely the finest food and drink. (Coppock and Beaver, 2013, ex. 178)

Capitalizing on whether (or how robustly) different exclusives allow both readings would allow us to make predictions about how much inference rates should increase in an experiment such as

Experiment 13. If, for instance, we were to conduct a manipulation with an exclusive that (unlike *only*) does not allow complement-exclusion readings, we would expect higher (and more uniform) rates of "Yes" responses.

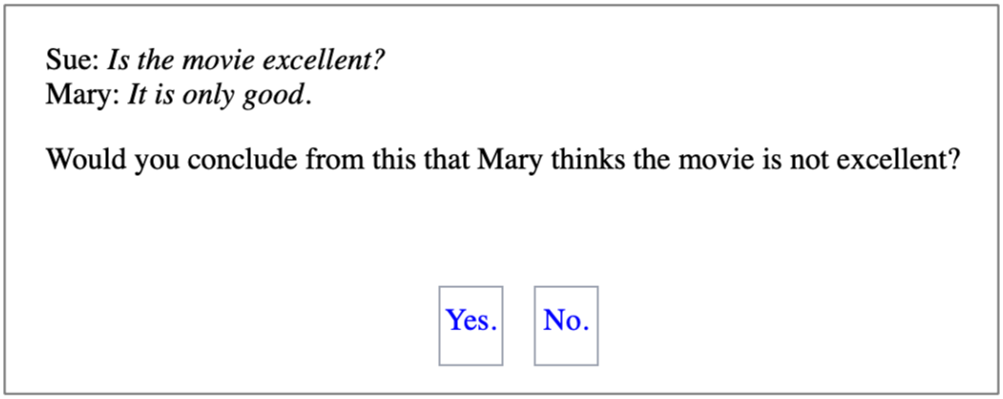
5.4 Experiment 14: Manipulating both cues

Experiments 12-13 showed that a pragmatic manipulation (supportive discourse context) and a semantic one (exhaustification with *only*) both increase inference rates and reduce variation across lexical scales. However, we also saw that with either manipulation, some of the variation still remains. Experiment 14 therefore combines context with *only*.

5.4.1 Participants and task

40 native speakers of American English participated in an online (Ibex and PCIbex) experiment for $2 compensation. Participant recruitment, screening, and compensation was identical to Experiment 7. Data from all 40 participants is reported below.

Experiment 14 combined the manipulation of Experiments 12 and 13: the potentially inference-triggering sentences included the focus particle *only*, and they were also preceded by a polar question that made reference to the stronger scalar alternative—see Figure 5.9 for an example. Otherwise, the task and materials were identical to previous experiments.



Sue: *Is the movie excellent?*
Mary: *It is only good.*

Would you conclude from this that Mary thinks the movie is not excellent?

Figure 5.9: Example experimental trial from Experiment 14

5.4.2 Hypothesis and predictions

The predictions made for Experiments 12 and 13 straightforwardly carry over to Experiment 14. First, because of Sue's explicit question, only on its inference-enriched meaning is Mary's answer congruent, which predicts increased inference rates and less variation across scales (see Section 5.2.2 for details). Second, *only* encodes the exclusion of alternatives in the semantics, which similarly predicts increased inference rates and less variation (Section 5.3.2).

Moreover, in Section 5.3.3 I argued that the reason Experiment 13's manipulation with *only* did not lead to uniformity and ceiling-level inference rates is that participants may have been excluding non-scalar alternatives. The discourse manipulation of Experiment 14 provides a clear alternative to be excluded, and I therefore predict that the variation that remained in Experiment 13 should be eliminated.

5.4.3 Results and discussion

Figure 5.10 shows the results of Experiment 14 (rightmost facet: "QUD + only"), along with all relevant previous experiments. A statistical analysis identical to the one reported in Section 5.2.3 confirms that Experiment 14's manipulation significantly increased rates of inference calculation as compared to Experiment 7's baseline (Estimate=3.74, SE= 0.35, z=10.64, p <0.001). As can be seen in the figure, inference rates are now in fact almost at ceiling.

The relative entropy resulting from Experiment 14 is 0.006—see Table 5.3 for a comparison of the relative entropy from all experiments. We can see that uniformity was very nearly achieved in Experiment 14; we no longer find appreciable variation in inference rates across the lexical scales tested.

Overall, Experiments 12–14 are informative as to the interplay of different factors that can make the calculation of upper-bounded inferences more likely and more uniform. We can interpret these experiments as manipulating two such factors: first, whether the specific alternative is made salient in the discourse context; and second, the degree to which the linguistic expression encourages alternative exclusion. We saw that in Experiment 12 scalar diversity was reduced, but

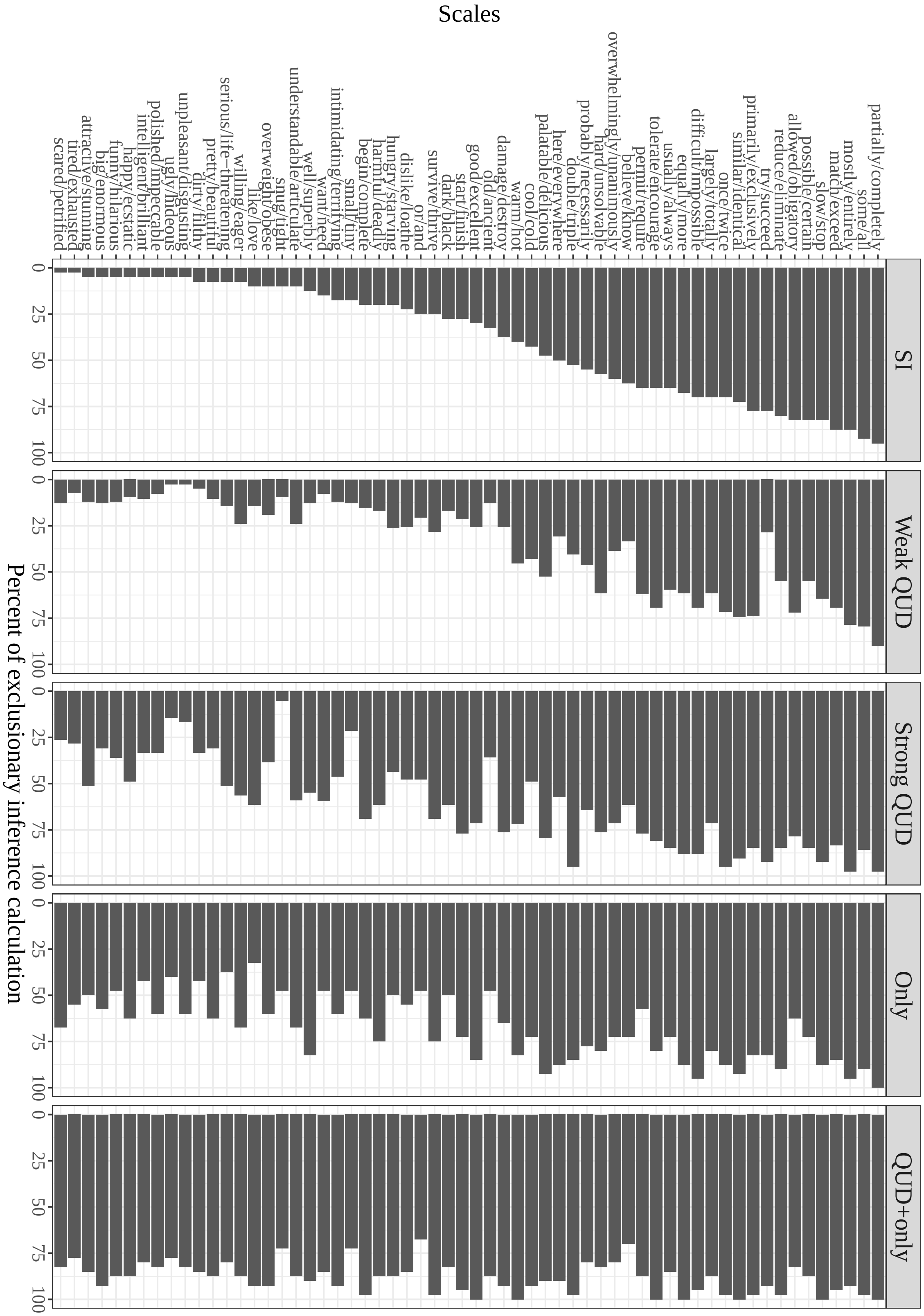


Figure 5.10: SI rate for 60 different scales. Experiments 7 (SI), 12 (Weak QUD, Strong QUD), 13 (Only) and 14 (QUD + only) are shown shown on the five facets of the plot.

	Manipulation	Relative entropy
Experiment 7	Baseline scalar diversity	0.466
Experiment 12	Weak QUD	0.378
Experiment 12	Strong QUD	0.123
Experiment 13	Exhaustification with *only*	0.046
Experiment 14	Context and *only*	0.006

Table 5.3: Relative entropy results by experiment: Experiments 7, 12, 13 and 14

not eliminated. This is because discourse context can provide salient alternatives, but that alone does not tell hearers that they need to reason about and exclude those alternatives. Properties of the linguistic expression, however, do: the focus particle *only* makes reasoning about and excluding alternatives obligatory. However, the presence of *only* does not make it clear what the relevant alternatives are. This is why variation still remained in Experiment 13. As Experiment 14 demonstrates, only when both of these factors are fixed – the identity of the alternatives is made clear, and the cue to exclude them is encoded semantically – do we find ceiling effects and uniformity in inference calculation. When either of these supportive cues is absent, there is more flexibility in interpretation, and consequently we observe more variation. This also leaves more opportunity for other factors (reviewed and investigated in Chapter 4, e.g., distinctness, extremeness, etc.) to influence the likelihood of inference calculation.

5.5 Summary of findings

In this chapter I investigated what factors can increase the rate of exclusionary inference calculation and introduce uniformity. My findings revealed that a pragmatic manipulation (explicit question) and a semantic manipulation (exhaustification with *only*) both lead to increased inference rates and reduced diversity—the latter more so than the former. However, we saw that variation still remains under either manipulation, and only when we combine them do we find ceiling level inference rates and uniformity across scales. This suggests an important role of both discourse context and alternative exclusion in the likelihood of and variation in inference calculation.

CHAPTER 6

CONCLUSION

This dissertation investigated pragmatic phenomena where hearers regularly infer meanings stronger than what was literally said. Examples (1) and (2) are instances of scalar inference —called scalar, because <*some, all*> and <*good, excellent*> are taken to form scales.

(1) Mary ate some of the deep dish.
SI: Mary didn't eat all of the deep dish.

(2) The movie is good.
SI: The movie isn't excellent.

Alternatives like *all* and *excellent* are informationally stronger than their weaker scale-mates, *some* and *good*. An utterance containing the weaker terms, then, can be taken to imply the negation of the stronger alternative. If the speaker had been in a position to utter *The movie is excellent*, she should have done so in order to maximize informativity. But she chose not to utter the stronger alternative, leading hearers to infer its negation.

Another crucial property of scalar inference is context-sensitivity. In the context of a question like (3-a), (1) is more likely to lead to the *some but not all* SI than in the context of a question like (3-b). And similarly, the *good but not excellent* SI is more likely following a question like (4-a), as compared to (4-b).

(3) a. Did Mary eat all of the deep dish?
b. Did Mary eat any of the deep dish?

(4) a. Is the movie excellent?
b. Is the movie good?

In this dissertation, I investigated the respective roles of lexical alternatives and discourse context in the calculation and processing of SI. In Chapter 2, we saw evidence that lexical alternatives (*all*, *excellent*) are activated in the processing of SI-triggering sentences, providing evidence for the psychological reality of scales. In Chapter 3, on the other hand, we saw that the processing cost of SI calculation, as evidenced by increased reaction times, is better explained by whether the discourse context supports or discourages SI calculation, than by whether lexical alternatives were retrieved. In Chapter 4, I sought explanations of scalar diversity —e.g., the finding that the SI in (1) arises more robustly than the one in (2) —by looking at various properties of the stronger alternatives. We saw that some (but not all) properties of different alternatives can indeed predict the likelihood of SI calculation across scales, but overall most of scalar diversity cannot be reduced to differences across alternatives. Lastly, in Chapter 5, I investigated the effect of a supportive discourse context vs. grammatically encoding the exclusion of alternatives (using the focus particle *only*) on the rate and uniformity of inference calculation. We saw that only when both cues align does inference calculation become deterministic.

Altogether, the findings of this dissertation suggest that global properties of the discourse context and local properties of the scalar alternatives themselves both play an important role in some (but not all) aspects of SI calculation and processing.

REFERENCES

Alexandropoulou, Stavroula. 2022. Scalar diversity and ignorance inferences: An experimental study on *at least* as a modifier of numerals vs. adjectives. In *Proceedings of Semantics and Linguistic Theory (SALT) 31*, ed. Nicole Dreier, Chloe Kwon, Thomas Darnell, and John Starr, 506–529.

Atlas, Jay D., and Stephen C. Levinson. 1981. It-clefts, informativeness and logical form. In *Radical pragmatics*, ed. P. Cole, 1–62. New York: Academic Press.

Baker, Rachel, Ryan Doran, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1:211–248.

Banga, Arina, Ingeborg Heutinck, Sanne M. Berends, and Petra Hendriks. 2009. Some implicatures reveal semantic differences. *Linguistics in the Netherlands* 26:1–13.

Barner, David, Neon Brooks, and Alan Bale. 2011. Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118:84–93.

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Beaver, David I., and Brady Z. Clark. 2008. *Sense and sensitivity*. Wiley-Blackwell.

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

Birch, S.L., and S.M. Garnsey. 1995. The effect of focus on memory for words in sentences. *Journal of Memory and Language* 34:232–267.

Bonnefon, Jean-François, Aidan Feeney, and Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112:249–258.

Bott, Lewis, Todd M. Bailey, and Daniel Grodner. 2012. Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language* 66:123–142.

Bott, Lewis, and Emmanuel Chemla. 2016. Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language* 91:117–140. New Approaches to Structural Priming.

Bott, Lewis, and Steven Frisson. 2022. Salient alternatives facilitate implicatures. *PLOS ONE* 17:1–10.

Bott, Lewis, and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51:437–457.

Braun, Bettina, and Lara Tagliapietra. 2009. The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes* 25:1024–1043.

Breheny, Richard. 2019. Scalar implicatures. In *The oxford handbook of experimental semantics and pragmatics*, ed. Chris Cummins and Napoleon Katsos, 38–61. OUP Oxford.

Breheny, Richard, Napoleon Katsos, and John Williams. 2006. Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100:434–463.

de Carvalho, Alex, Anne C. Reboul, Jean-Baptiste Van der Henst, Anne Cheylus, and Tatjana Nazir. 2016. Scalar implicatures: The psychological reality of scales. *Frontiers in Psychology* 7.

Chemla, Emmanuel, and Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130:380–396.

Chemla, Emmanuel, and Benjamin Spector. 2011. Experimental Evidence for Embedded Scalar Implicatures. *Journal of Semantics* 28:359–400.

Chevallier, Coralie, Ira A. Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti, and Dan Sperber. 2008. Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology* 61:1741–1760. PMID: 18942038.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In *Structures and beyond*, ed. Adriana Belletti, 39–103. Oxford University Press.

Chierchia, Gennaro, Stephen Crain, Maria Teresa Guasti, Andrea Gualmini, and Luisa Meroni. 2001. The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *BUCLD 25 Proceedings*, ed. Aimee Johansen Anna H.-J. Do, Laura Domínguez. Somerville, Massachusetts: Cascadilla Press.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In *Semantics: An international handbook of natural language meaning*, ed. Klaus von Heusinger, Claudia Maienborn, and Paul Portner, volume 3, 2297–2331. Mouton de Gruyter.

Chomsky, Noam. 1972. *Studies on Semantics in Generative Grammar*. The Hague: Mouton.

Clifton, Charles Jr, and Chad Dube. 2010. Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics* 3:1–13.

Constant, Noah. 2012. English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy* 35:407–442.

Coppock, Elizabeth, and David I. Beaver. 2013. Principles of the Exclusive Muddle. *Journal of Semantics* 31:371–432.

Cummins, Chris, and Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology* 6:1779.

Davies, Mark. 2008. The corpus of contemporary american english (coca). https://www.english-corpora.org/coca/.

De Neys, Wim, and Walter Schaeken. 2007. When people are more logical under cognitive load – Dual task impact on scalar implicature. *Experimental Psychology* 54:128–133.

Degen, Judith. 2013. Alternatives in Pragmatic Reasoning. Doctoral Dissertation, University of Rochester.

Degen, Judith, and Michael K. Tanenhaus. 2014. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39:667–710.

Degen, Judith, and Michael K. Tanenhaus. 2019. Constraint-based pragmatic processing. In *Handbook of experimental semantics and pragmatics*, ed. Chris Cummins and Napoleon Katsos. Oxford: Oxford University Press.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88:124–154.

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

Feeney, Aidan, Susan Scrafton, Amber Duckworth, and Simon J Handley. 2004. The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology* 58:121–132.

Ferguson, Heather J., Anthony J. Sanford, and Hartmut Leuthold. 2008. Eye-movements and erps reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research* 1236:113–125.

von Fintel, Kai, and Danny Fox. 2002. Classnotes for 24:954: Pragmatics in linguistic theory. https://dspace.mit.edu/handle/1721.1/36355.

Fox, Danny, and Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19:87–107.

Fraundorf, Scott H., Aaron S. Benjamin, and Duane G. Watson. 2013. What happened (and what did not): Discourse constraints on encoding of plausible alternatives. *Journal of Memory and Language* 69:196–227.

Fraundorf, Scott H., Duane G. Watson, and Aaron S. Benjamin. 2010. Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language* 63:367–386.

Gazdar, Gerald. 1977. *Implicature, presupposition, and logical form.* Bloomington: Indiana University Linguistics Club.

Geurts, Bart. 2010. *Quantity implicatures*. Cambridge University Press.

Geurts, Bart, and Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics* 2:1–34.

Geurts, Bart, and Bob van Tiel. 2013. Embedded scalars. *Semantics and Pragmatics* 6:1–37.

Goodman, Noah D., and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5:173–184.

Gotzner, Nicole. 2017. *Alternative sets in language processing: How focus alternatives are represented in the mind*. Palgrave Macmillan.

Gotzner, Nicole, and Jacopo Romoli. 2022. Meaning and alternatives. *Annual Review of Linguistics* 8:213–234.

Gotzner, Nicole, Stephanie Solt, and Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9:1659.

Gotzner, Nicole, and Katharina Spalek. 2017. Role of contrastive and noncontrastive associates in the interpretation of focus particles. *Discourse Processes* 54:638–654.

Gotzner, Nicole, and Katharina Spalek. 2019. The life and times of focus alternatives: Tracing the activation of alternatives to a focused constituent in language comprehension. *Language and Linguistics Compass* 13:e12310.

Gotzner, Nicole, Isabell Wartenburger, and Katharina Spalek. 2016. The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition* 8:59–95.

Green, Peter, and Catriona J. MacLeod. 2016. SIMR: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* 7:493–498.

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

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116:42–55.

Groenendijk, Jeroen, and Martin Stokhof. 1984. On the semantics of questions and the pragmatics of answers. In *Varieties of Formal Semantics: Proceedings of the Fourth Amsterdam Colloquium*, ed. Fred Landman and Frank Veltman, 143–170. Foris.

Gualmini, Andrea, Sarah Hulsey, Valentine Hacquard, and Danny Fox. 2008. The Question-Answer Requirement for scope assignment. *Natural Language Semantics* 16:205–237.

Guasti, Maria Teresa, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini, and Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20:667–696.

Göbel, Alex, and Michael Wagner. 2022. On a concessive reading of the rise-fall-rise contour: contextual and semantic factors. Talk presented at Experiments in Linguistic Meaning 2.

Hale, John. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research* 32:101–123.

Hamblin, Charles L. 1976. Questions in montague english. In *Montague grammar*, ed. Barbara H. Partee, 247–259. New York: Academic Press.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10:146–162.

Hawkins, Robert X. D., Andreas Stuhlmüller, Judith Degen, and Noah D. Goodman. 2015. Why do you ask? Good questions provoke informative answers. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, ed. David C. Noelle, Rick Dale, Anne Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn Jennings, and Paul P. Maglio, 878–883. Austin, TX: Cognitive Science Society.

Hirschberg, Julia Bell. 1985. A theory of scalar implicature. Doctoral Dissertation, University of Pennsylvania.

Horn, Laurence R. 1969. A presuppositional analysis of *only* and *even*. In *The Fifth Regional Meeting of the Chicago Linguistics Society (CLS 5)*, 98–107.

Horn, Laurence R. 1972. On the semantic properties of logical operators in English. Doctoral Dissertation, UCLA.

Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In *Meaning, form, and use in context: Linguistic applications*, ed. Deborah Schiffrin, 11–42. Washington D.C.: Georgetown University Press.

Horn, Laurence R. 2000a. From if to iff: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics* 32:289–326.

Horn, Laurence R. 2000b. Pick a theory (not just any theory): Indiscriminatives and the free-choice indefinite. In *Negation and polarity: Syntactic and semantic perspectives*, ed. Laurence R. Horn and Yasuhiko Kato, 147–192. Oxford, UK: Oxford University Press.

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

Huang, Yi Ting, and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology* 58:376–415.

Hulsey, Sarah, Valentine Hacquard, Danny Fox, and Andrea Gualmini. 2004. The Question-Answer Requirement and scope assignment. In *MIT Working Papers in Linguistics*, ed. Aniko Csirmaz, Andrea Gualmini, and Andrew Nevins, 71–90. MITWPL.

Husband, E. Matthew, and Fernanda Ferreira. 2015. The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience* 31:217–235.

Jackendoff, Ray S. 1974. *Semantic Interpretation in Generative Grammar (current Studies in Linguistics)*. MIT Press.

Jurafsky, Dan, and James H. Martin. 2020. Speech and Language Processing. 3rd ed. draft.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30:669–690.

Kaup, Barbara, and Rolf A. Zwaan. 2003. Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:439–446.

Kaup, Barbara, Rolf A. Zwaan, and Jana Lüdtke. 2007. The experiential view of language comprehension: How is negation represented? In *Higher level language processes in the brain: Inference and comprehension processes*, ed. Franz Schmalhofer and Charles A. Perfetti, 255–288. Mahwah, New Jersey and London: Lawrence Erlbaum Associates.

Kennedy, Christopher, and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81:345–381.

Kim, Christina S., Christine Gunlogson, Michael K. Tanenhaus, and Jeffrey T. Runner. 2015. Context-driven expectations about focus alternatives. *Cognition* 139:28–49.

Kroch, Anthony. 1972. Lexical and inferred meanings for some time adverbials. *Quarterly Progress Reports of the Research Laboratory of Electronics* 104:260–267.

Kullback, Solomon, and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.

Kursat, Leyla, and Judith Degen. 2020. Probability and processing speed of scalar inferences is context-dependent. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, ed. Stephanie Denison, Michael Mack, Yang Xu, and Blair C. Armstrong, 1236–1242. Cognitive Science Society.

Kutas, Marta, and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307:161–163.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82:1–26.

Landauer, Thomas K, and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104:211–240.

Levinson, Stephen C. 2000. *Presumptive Meanings*. MIT Press Ltd.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106:1126–1177.

Lupker, Stephen J., and Penny M. Pexman. 2010. Making things difficult in lexical decision: the impact of pseudohomophones and transposed-letter nonwords on frequency and semantic priming effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36:1267–1289.

de Marneffe, Marie-Catherine, and Judith Tonhauser. 2019. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In *Current research in the semantics/pragmatics interface*, ed. Malte Zimmermann, Klaus von Heusinger, and Edgar Onea, volume 36, Questions in Discourse, 132–163. Leiden, The Netherlands: Brill.

Matsumoto, Yo. 1995. The conversational condition on Horn scales. *Linguistics and Philosophy* 18:21–60.

Mikolov, Tomas, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. International Conference on Learning Representations, http://arxiv.org/abs/1301.3781.

Musolino, Julien. 1998. Universal grammar and the acquisition of semantic knowledge: An experimental investigation of quantifier–negation interaction in English. Doctoral Dissertation, University of Maryland, College Park.

Musolino, Julien. 2011. Studying language acquisition through the prism of isomorphism. In *Handbook of generative approaches to language acquisition*, ed. Jill de Villiers and Tom Roeper, 319–349. Dordrecht: Springer Netherlands.

Noveck, Ira. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78:165–188.

Noveck, Ira A., Gennaro Chierchia, Florelle Chevaux, Raphaelle Guelminger, and Emmanuel Sylvestre. 2002. Linguistic-pragmatic factors in interpreting disjunctions. *Thinking & Reasoning* 8:297–326.

Noveck, Ira A., and Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85:203–210.

Pankratz, Elizabeth, and Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13:562–594.

Papafragou, Anna, and Julien Musolino. 2003. Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition* 86:253–282.

Papafragou, Anna, and Niki Tantalou. 2004. Children's computation of implicatures. *Language Acquisition* 12:71–82.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pijnacker, Judith, Peter Hagoort, Jan Buitelaar, Jan-Pieter Teunisse, and Bart Geurts. 2009. Pragmatic inferences in high-functioning adults with autism and asperger syndrome. *Journal of Autism and Developmental Disorders* 39:607–618.

Politzer-Ahles, Stephen, and Robert Fiorentino. 2013. The realization of scalar inferences: Context sensitivity without processing cost. *PLOS ONE* 8:1–6.

Pouscoulous, Nausicaa, Ira A. Noveck, Guy Politzer, and Anne Bastide. 2007. A developmental investigation of processing costs in implicature production. *Language Acquisition* 14:347–375.

Rastle, Kathleen, Jonathan Harrington, and Max Coltheart. 2002. 358,534 nonwords: The arc nonword database. *The Quarterly Journal of Experimental Psychology Section A* 55:1339–1362. PMID: 12420998.

Rees, Alice, and Lewis Bott. 2018. The role of alternative salience in the derivation of scalar implicatures. *Cognition* 176:1–14.

Řehůřek, Radim, and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Repp, Sophie, and Katharina Spalek. 2021. The role of alternatives in language. *Frontiers in Communication* 6.

Roberts, Craige. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5:1–69.

Ronai, Eszter, and Ming Xiang. 2021. Exploring the connection between Question Under Discussion and scalar diversity. *Proceedings of the Linguistic Society of America* 6:649–662.

Rooth, Mats. 1985. Association with focus. Doctoral Dissertation, University of Massachusetts, Amherst, Amherst.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1:75–116.

Sanford, Alison J. S., Jessica Price, and Anthony J. Sanford. 2009. Enhancement and suppression effects resulting from information structuring in sentences. *Memory & Cognition* 37:880–888.

Schwarz, Florian, Zehr Jérémy, Grodner Daniel, and Hezekiah Akiva Bacovcin. 2016. Subliminal priming of alternatives does not increase implicature responses. Poster presented at the Logic and Language in Conversation Workshop, University of Utrecht.

Simons, Mandy, and Tessa Warren. 2018. A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology* 71:272–279.

Singh, Raj, Ken Wexler, Andrea Astle-Rahim, Deepthi Kamawar, and Danny Fox. 2016. Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics* 24:305–352.

Spalek, Katharina, Nicole Gotzner, and Isabell Wartenburger. 2014. Not only the apples: Focus sensitive particles improve memory for information-structural alternatives. *Journal of Memory and Language* 70:68–84.

Sperber, Dan, and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. Wiley-Blackwell, 2nd edition edition.

Storto, Gianluca, and Michael K Tanenhaus. 2005. Are scalar implicatures computed online? In *Proceedings of Sinn und Bedeutung*, volume 9, 431–445.

Sturt, Patrick, Anthony J. Sanford, Andrew Stewart, and Eugene Dawydiak. 2004. Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review* 11:882–888.

Sun, Chao, and Richard Breheny. 2021. What the inference task can tell us about the comprehension of scalars and numbers: An investigation of probe question and response bias. Talk presented at Sinn und Bedeutung 26, https://osf.io/6xmwr.

Sun, Chao, Ye Tian, and Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9.

Swinney, David A. 1979. Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18:645–659.

Swinney, David A, William Onifer, Penny Prather, and Max Hirshkowitz. 1979. Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory & Cognition* 7:159–165.

Taylor, Wilson L. 1953. "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly* 30:415–433.

Thomas, Matthew A., James H. Neely, and Patrick O'Connor. 2012. When word identification gets tough, retrospective semantic processing comes to the rescue. *Journal of Memory and Language* 66:623–643.

Tian, Ye, Heather Ferguson, and Richard Breheny. 2016. Processing negation without context–why and when we represent the positive argument. *Language, Cognition and Neuroscience* 31:683–698.

van Tiel, Bob. 2013. Embedded Scalars and Typicality. *Journal of Semantics* 31:147–177.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33:137–175.

Van Kuppevelt, Jan. 1996. Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19:393–443.

van Tiel, Bob, and Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science* 41:1119–1154.

de Vega, Manuel, and Mabel Urrutia. 2012. Discourse updating after reading a counterfactual event. *Psicologica: International Journal of Methodology and Experimental Psychology* 33:157–173.

Ward, Gregory, and Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language* 61:747–776.

Westera, Matthijs, and Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung* 24:439–454.

Yan, Mengzhu, and Sasha Calhoun. 2019. Priming effects of focus in mandarin chinese. *Frontiers in Psychology* 10.

Yang, Xiao, Utako Minai, and Robert Fiorentino. 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology* 9:1720.

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

Zhang, Dabao. 2021. R-Squared and Related Measures. https://cran.r-project.org/package=rsq.

Zondervan, Arjen. 2010. Scalar implicatures or focus: an experimental approach. Doctoral Dissertation, Universiteit Utrecht.

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. Tova Friedman and Satoshi Ito, 765–777.